

Recognising Biomedical Names: Challenges and Solutions

XIANG DAI



THE UNIVERSITY OF
SYDNEY

Supervisors: Sarvnaz Karimi
Ben Hachey
Cecile Paris
Joachim Gudmundsson

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

2021

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Xiang Dai

2021-Feb-27

Abstract

The growth rate in the amount of biomedical documents—such as scholarly articles, clinical notes and health forum discussions—is staggering. Unlocking information trapped in these documents can enable researchers and practitioners to operate confidently in the information world. Biomedical Information Extraction (IE) system aims to automatically extract structured information—such as biomedical concepts, attributes, events, and their relations—from unstructured text. Within an IE system, the first step is called Biomedical Named Entity Recognition (NER), the task of recognising biomedical names.

NER has been heavily studied in the generic domain, recognising person, organisation, and location names in newspaper articles. However, the effectiveness of existing Biomedical NER model is still not satisfactory. In contrast to entity mentions in the generic domain which are usually short spans of text, biomedical names—surface forms that represent biomedical concepts, such as genes, proteins, symptoms, diseases, and drugs—pose unique challenges. For example, it is even common for an ordinary person to confuse ‘*severe acute respiratory syndrome coronavirus 2*’ (virus name), ‘*severe acute respiratory syndrome*’ (disease name), and ‘*coronavirus disease 2019*’ (disease name). The variety of language used for different communicative purposes makes biomedical NER even more challenging. Various groups of people use totally different languages to describe the same biomedical concept. For example, researchers tend to use standard names in biomedical vocabularies to make the description more comprehensible and less confused; hospital doctors, who write notes under time pressure, use abbreviations for efficient communication with their colleagues; and, ordinary people use linguistically noisy layman language to share their experiences.

State-of-the-art NER models, based on sequence tagging technique, are good at recognising short entity mentions in the generic domain, especially when they are enhanced by pre-trained language representation models. However, there are several open challenges of applying these models to recognise biomedical names:

- Biomedical names may contain complex inner structure (discontinuity and overlapping) which cannot be recognised using standard sequence tagging technique;
- The training of NER models usually requires large amount of labelled data, which are difficult to obtain in the biomedical domain; and,
- Commonly used language representation models are pre-trained on generic data, such as the Wikipedia and books, a domain shift therefore exists between these models and target biomedical data.

To deal with these challenges, we explore several research directions and make the following contributions: (1) we propose a transition-based NER model which can recognise discontinuous mentions. Through experiments on three datasets from the biomedical domain, we show that our model can effectively recognise discontinuous entity mentions without sacrificing the accuracy on continuous mentions. Analysis also suggests that our model is good at recognising long mentions, resulting in higher recall than other baselines; (2) We develop a cost-effective approach that nominates the suitable pre-training data, via measuring the similarity between different pre-training data options and target task data. Through experiments on 56 source-target data pairs, we show that simple similarity measures are good predictors of the usefulness of pre-trained language representation models on downstream NER datasets; and, (3) We design several data augmentation methods which do not rely on any external trained models, for NER. Experimental results show that the proposed augmentation methods can improve performance over strong baselines, where large scale pre-trained language representation models are used.

Our contributions have obvious practical implications, especially when new biomedical applications are needed. Our proposed data augmentation methods can help the NER model achieve decent performance, requiring only a small amount of labelled data. Our investigation regarding selecting pre-training data can improve the model by incorporating language representation models, which are pre-trained using in-domain data. Finally, our proposed transition-based NER model can further improve the performance by recognising discontinuous mentions without sacrificing the accuracy on continuous mentions.

Acknowledgements

I enjoyed my journey of doing a PhD, and I am immensely thankful for the many people I met during this journey. Foremost thanks must go to my PhD supervisors Sarvnaz Karimi, Ben Hachey and Cecile Paris. Sarvnaz, thank you for your encouragement, without which I would never have started working on NLP. Also thank you for your detailed guidance, which shapes my research. Ben, thank you for your completely honest criticisms and feedback, which always help me revisit my work from a practical perspective. Cecile, thank you for sensible advice and patience. I learn a lot from people like you who want to make things to be perfect. I also want to thank my supervisor Joachim Gudmundsson, who helped me a lot with university's administration and funding.

I am very thankful to Dietrich Klakow from Saarland University. Because of the COVID-19 pandemic, I was stranded in Germany after I finished my internship at Bosch Center for Artificial Intelligence. Dietrich hosted me in his group, providing me a shelter where I could finish my thesis. Thanks also to Heike Adel, Matthew Honnibal, and Vera Demberg for their help.

I also want to thank my colleagues at Data61. In particular, I am grateful to Aditya Joshi, Chang Xu, Maciej Rybinski, Vincent Nguyen, Stephen Wan, Sunghwan Mac Kim, Wenyi Tay, and Sonit Singh for regular reading group meetings, and inspiring discussions. Thanks also to Lukas Lange and Michael A. Hedderich from Saarland University for feedback on drafts of this thesis, and exchange of ideas.

Finally, thank you to my family for being encouraging and patient. Special thanks to my dad, who explains me a lot about medications and medical procedures.

Authorship Attribution Statement

- The Chapter 1 of this thesis relates to (Dai et al., 2017).

Xiang Dai, Sarvnaz Karimi, and Cecile Paris. 2017. Medication and adverse event extraction from noisy text. In Proceedings of the Australasian Language Technology Association Workshop, pages 79–87, Brisbane, Australia.
- The Section 2.4 of this thesis relates to (Dai, 2018).

Xiang Dai. 2018. Recognizing complex entity mentions: A review and future directions. Proceedings of ACL 2018, Student Research Workshop, pages 37–44, Melbourne, Australia.
- Chapter 3 of this thesis relates to (Dai and Adel, 2020).

Xiang Dai, Heike Adel. 2020. An Analysis of Simple Data Augmentation for Named Entity Recognition. In Proceedings of the 28th International Conference on Computational Linguistics, Online.
- Chapter 4 of this thesis relates to (Dai et al., 2019, 2020a).

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2019. Using similarity measures to select pretraining data for NER. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1460–1470, Minneapolis, Minnesota.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. Cost-effective Selection of Pretraining Data: A Case Study of Pretraining BERT on Social Media. In Findings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online.
- Chapter 5 of this thesis relates to (Dai et al., 2020b).

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An effective transition-based model for discontinuous NER. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5860–5870, Online.

I contribute to every aspect of the above mentioned publications, including designing the method, conducting the experiment, and writing the paper.

Contents

Statement of Originality	ii
Abstract	iii
Acknowledgements	v
Authorship Attribution Statement	vi
Contents	viii
List of Figures	xii
List of Tables	xvi
Notations	1
Abbreviations	2
Chapter 1 Introduction	4
1.1 Recognising Adverse Drug Events from Social Media—A Motivating Application	5
1.1.1 A unified architecture for sequence taggers	8
1.2 Key Open Challenges	14
1.2.1 Biomedical names are complex	14
1.2.2 Labelling data is difficult	14
1.2.3 Unlabelled biomedical data are limited	16
1.3 About the Thesis	18
1.3.1 Publications	19
1.3.2 Definition and clarification of used terms	20
1.4 Summary	22

Chapter 2 Literature Review	24
2.1 A Brief History of NER	24
2.1.1 Dictionary based approach	26
2.1.2 Rule based approach	27
2.1.3 Statistical machine learning based approach	29
2.2 NLP for Low Resource Scenarios	30
2.2.1 Distant supervision	31
2.2.2 Active learning	34
2.2.3 Data augmentation	37
2.2.4 Summary	41
2.3 Transfer Learning	42
2.3.1 Cross-task transfer	43
2.3.2 Cross-domain transfer	47
2.3.3 Summary	49
2.4 Complex Entity Recognition	50
2.4.1 Definitions of complex entity mentions	51
2.4.2 Token-level Approach	53
2.4.3 Span-based Approach	55
2.4.4 Sentence-level Approach	59
2.4.5 Summary	63
Chapter 3 Data Augmentation for NER	65
3.1 Overview	65
3.2 Proposed Data Augmentation Methods	66
3.3 Evaluation	69
3.3.1 Datasets	69
3.3.2 Backbone model	70
3.3.3 Experimental results	72
3.4 Analysis	73
3.4.1 The impact of hyperparameters	73
3.4.2 A closer look at errors	76

3.5	Summary	77
Chapter 4 Cost-effective Selection of Pre-training Data		78
4.1	Overview	78
4.2	What Human Intuition Indicates	80
4.3	Similarity Measures	82
4.3.1	Target vocabulary covered	82
4.3.2	Jaccard similarity of vocabularies	83
4.3.3	Language model perplexity	83
4.3.4	Jensen-Shannon divergence	84
4.4	Datasets	85
4.5	Experimental Results	86
4.5.1	Predictiveness of similarity measures	91
4.5.2	Comparison to publicly available pre-trained models	92
4.6	Summary	93
Chapter 5 Transition-based Model for Discontinuous NER		94
5.1	Overview	94
5.2	Datasets	96
5.3	Proposed Model	99
5.3.1	Representation of the parser state	100
5.3.2	Capturing discontinuous dependencies	102
5.3.3	Selecting an action	102
5.4	Experimental Results	103
5.4.1	Baseline models	103
5.4.2	Experimental setup	104
5.4.3	Results	104
5.5	Analysis	107
5.5.1	Impact of mention and interval length	107
5.5.2	Impact of overlapping structure	108
5.5.3	Example predictions	110

5.5.4 Ablation studies	111
5.6 Summary	112
Chapter 6 Conclusions	113
Bibliography	116

List of Figures

- 1.1 An example input sentence and the entity mentions which are supposed to be recognised by the NER model. 7
- 1.2 State-of-the-art NER model is based on sequence tagging technique that assigns a tag to each token. Token positions of mentions can be extracted from the output tag sequence. 7
- 1.3 A unified architecture for sequence tagging models, consisting of a mapping function and a classifier. 8
- 1.4 An LSTM computes the output state by taking the entire past (left context) of the input sequence into consideration. 10
- 1.5 An forward model computes the output state by taking the entire past (left context) of the input sequence into consideration, whereas the backward model considers the entire right context. Both output hidden states are concatenated to form the final contextual string embedding and capture the information of the token itself as well as its surrounding tokens. 12
- 1.6 The semi-supervised learning approach used in FLAIR: pre-training two—forward and backward—character level language models, and using the pre-trained model as part of the mapping function in the downstream supervised tasks. For the sake of brevity, we show only the forward model and the backward model is omitted. The whitespace character is represented using " " in this figure. 13
- 1.7 We focus on Biomedical NER, and explore the following three research directions: recognising discontinuous entity mentions; pre-training domain-specific language representation models; and enhancing the effectiveness of NER models using data augmentation. 18

- 2.1 Example sequence of tags generated by a rule and two domain-specific dictionaries. For example, ‘*Cymbalta*’ is assigned the tag ‘*I-Drug*’ because it appears in the drug dictionary. 32
- 2.2 The active learning process usually have multiple rounds, each of which consists of five steps: (1) applying model on unlabelled data; (2) querying on unlabelled data; (3) presenting informative instances; (4) annotating instances; and (5) updating the model. 35
- 2.3 The extractive question answering model tends to use the question type (e.g., Who) and select the spans whose nature agrees with the question type (e.g., ‘Bill Clinton’, ‘George H. W. Bush’, and ‘Ross Perot’), without the necessity to understand the question. 39
- 2.4 Neural architectures for the settings of cross-domain, cross-task, cross-lingual transfer proposed in (Yang et al., 2017). 44
- 2.5 The Skip-gram model aims to learn word representations that can be used to predict the surrounding words. 45
- 2.6 The mask language modelling pre-training task aims to learn contextual word representations that can be used to predict what the masked token is. 47
- 2.7 The replace token detection task aims to train the discriminator to predict whether the token is the original token or a fictional token. 48
- 2.8 Examples involving nested, overlapping and discontinuous entity mentions. In (a), ‘*HIV-1 enhancer*’ and ‘*HIV-1*’ are nested entity mentions. In (b), ‘*intense pelvic pain*’ and ‘*back pain*’ overlap, and ‘*intense pelvic pain*’ is a discontinuous mention. 51
- 2.9 In a linear-chain CRF model, the output for each token depends on the representation of that token in context and the output for the previous token. 53
- 2.10 An encoding example of two adverse drug event mentions: ‘*intense pelvic pain*’ and ‘*back pain*’. 54
- 2.11 An example of mention separators encoding two nested entity mentions: ‘*IL2*’ and ‘*IL2 regulatory region*’. Muis and Lu (2017) design three mention separators: S, also denoted as \llbracket , indicating a mention is starting at the next token; E (\lrcorner), indicating a mention is ending at the previous token; and C ($-$), indicating a mention is continuing

- to the next token. X means none of the three separators applies. The standard sequence tagger, which takes as input a sequence of N tokens and outputs a sequence of $N-1$ mention separators, can be used to recognise nested NER. 55
- 2.12 An example of a sentence with three entity mentions: ‘*Bill Clinton*’ and ‘*Hilary Clinton*’ are PERSON mentions, and ‘*Canada*’ is a LOCATION mention. P and L refer to the entity categories: PERSON and LOCATION, respectively. 60
- 2.13 An example sub-hypergraph with two nested entity mentions: ‘*HIV-1*’ (VIRUS) and ‘*HIV-1 enhancer*’ (DNA). Here, one mention corresponds to a path consisting of (AETI+X) nodes. Specifically, the path $(A_4E_4T_4^1I_4^1X)$ corresponds to the mention ‘*HIV-1*’, and the path $(A_4E_4T_4^2I_4^2I_5^2X)$ corresponds to the mention ‘*HIV-1 enhancer*’. 62
- 2.14 An example sub-hypergraph with two entity mentions: ‘*muscle pain*’ and ‘*muscle fatigue*’. Muis and Lu (2016) extend the hypergraph representation proposed by Lu and Roth (2015) to capture discontinuous mentions through two new node types: B_k^i representing the k -th token is part of the i -th component of an entity mention, and O_k^i representing the k -th token appears in between $(i - 1)$ -th and i -th components of an entity mention. In this example, the path $(A_3E_3TB_3^0B_4^1X)$ corresponds to the mention ‘*muscle pain*’ and the path $(A_3E_3TB_3^0O_4^1O_5^1B_6^1X)$ corresponds to the discontinuous mention ‘*muscle fatigue*’. 63
- 3.1 High level overview of the BERT-CRF model. 70
- 3.2 Impact of the number of augmented instances per original training instance on the effectiveness of data augmentation. SR: synonym replacement. MR: mention replacement. 74
- 3.3 The impact of the ratio a token or a mention is replaced on the effectiveness of data augmentation. SR: synonym replacement. MR: mention replacement. 75
- 4.1 Survey questions regarding selection of pre-training data. 81
- 4.2 Likert scale ratings from NLP and ML practitioners ($N = 30$) for the statement ‘*Unsupervised pre-training on S would be useful for supervised named entity recognition learning on T.*’ Target data T is described as ‘*Online forum posts about medications,*’ source data S1 as ‘*Research papers about biology and health,*’ and source data S2 as ‘*Online reviews about restaurants, hotels, barbers, mechanics, etc.*’ 82

- 4.3 Correlation between different similarity measures and the effectiveness of domain-specific pre-trained models. 92
- 5.1 Examples involving discontinuous mentions, taken from the SHARE/CLEF 13 (Pradhan et al., 2013) and CADEC (Karimi et al., 2015a) datasets, respectively. The first example contains a discontinuous mention '*left atrium dilated*', the second example contains two mentions that overlap: '*muscle pain*' and '*muscle fatigue*' (discontinuous). 95
- 5.2 An example sequence of transitions. Given the states of stack and buffer (blue highlighted), as well as the previous actions, predict the next action (i.e., LEFT-REDUCE) which is then applied to change the states of stack and buffer. 101
- 5.3 The impact of mention length and interval length on recall. Mentions with interval length of zero are continuous mentions. Numbers in parentheses are the number of gold mentions. 108

List of Tables

1.1 The sequence of tokens is treated as a sequence of characters.	11
1.2 Discrepancy between the pre-trained model used in FLAIR and the target datasets: SHARE/CLEF 2013 (clinical notes) and CONLL 2003 (news stories).	17
2.1 The variants of word ‘ocular’ and the corresponding rules to generate them. The indentation reflects the hierarchical structure of these variants according to the history of how they are generated.	28
2.2 Decline in effectiveness of a model trained on NCBI-DISEASE (Doğan et al., 2014), when evaluated on other datasets: I2B2-2010 (Uzuner et al., 2011), and N2C2-2019 (n2c2, 2019). The mention-level F_1 score is reported.	30
2.3 Evaluation results, as reported by Lange et al. (2019), of automatically annotated labels against manual annotations.	32
2.4 Requirement of different types of resources by each approach.	42
2.5 A summary of techniques to represent and score span, given a sequence of token representations $\mathbf{h}_1, \dots, \mathbf{h}_n$. $\mathbf{h}(i, j)$, being a fixed-length vector representation of the span, with its dimension being a hyper-parameter. $\text{score}(i, j)$ is the (normalised) score for the span from i to j inclusive, where $1 \leq i \leq j \leq n$. $\text{score}(i, j)$ is usually a c -dimension vector, where c is the number of entity categories, including a special category for non-entity.	57
3.1 An original training instance and different types of augmented instances. We highlight changes using <i>italics</i> .	67
3.2 The descriptive statistics of the two English datasets from the biomedical domain: I2B2-2010 (Uzuner et al., 2011) and NCBI-DISEASE (Doğan et al., 2014).	69

3.3 Evaluation results in terms of span-level F_1 score. Small set contains 50 training instances; Medium contains 150 instances; Large contains 500 instances; Full uses the complete training set. Results that are better than the baseline model without using data augmentation are highlighted in bold. <u>underline</u> : the result is significantly better than the baseline model without data augmentation (paired student’s t-test, p: 0.05)	72
3.4 The comparison of different types of errors—FPs (false positives) and FNs (false negatives)—made by the baseline model without using data augmentation and models using Synonym Replacement (SR) and Mention Replacement (MR) data augmentation methods. \cap indicates the intersection of two sets, and \neg indicates the negative set. For example, the ‘FPs’ column corresponding to the ‘Baseline \cap SR \cap \neg MR’ row shows the number of false positives predicted by both the baseline model and the model using SR data augmentation, but not by the one using MR data augmentation.	76
4.1 Descriptive statistics of the source datasets.	86
4.2 List of the target NER datasets and their specifications.	87
4.3 Similarity values measured between source and target datasets. TVC: Target Vocabulary Covered. JSC: Jaccard similarity of Vocabularies. PPL: language model perplexity. JSD: Jensen-Shannon Divergence based on term distributions.	88
4.4 Similarity values measured between source and target datasets (continued). TVC: Target Vocabulary Covered. JSC: Jaccard similarity of Vocabularies. PPL: language model perplexity. JSD: Jensen-Shannon Divergence based on term distributions.	89
4.5 Pre-train hyper-parameters, which follow the practice of training ELECTRA-SMALL in (Clark et al., 2020).	91
4.6 The effectiveness of domain-specific pre-trained models on downstream NER tasks. We report the mention level F_1 scores.	91
4.7 Comparison between our best performing domain-specific models and the publicly available generic domain model.	93

5.1 The descriptive statistics of the datasets. ADE: adverse drug events; Disc.M: discontinuous mentions; Disc.M L.: discontinuous mention length, where intervals are not counted. Numbers in parentheses are the percentage of each category. Note that due to sentence segmentation issue, there are 13 and 64 mentions crossing multiple sentences in SHARE/CLEF 2013 and SHARE/CLEF 2014, respectively. We remove these mentions, as we frame the task as a sentence-level NER problem.	98
5.2 Evaluation results in terms of precision (P), recall (R) and F_1 score (F).	105
5.3 Evaluation results on sentences that contain at least one discontinuous mention.	106
5.4 Evaluation results on discontinuous mentions only.	106
5.5 Evaluation results on different categories of discontinuous mentions. ‘#’ columns show the number of gold discontinuous mentions in development set of each category.	109
5.6 Example sentences involving discontinuous entity mentions and predictions using different methods. These examples are taken from CADEC. Gold discontinuous mentions are highlighted in bold. We cross out the incorrect predictions (false positives) for easy understanding.	111
5.7 Ablation study to estimate the contribution of attention and ELMo components.	112

Notations

a	A scalar
\mathbf{a}	A vector
\mathbf{A}	A matrix
\mathbb{A}	A set
$P(\mathbf{a})$	A probability distribution over a discrete variable
$f(x; \boldsymbol{\theta})$	A function of x parameterised by $\boldsymbol{\theta}$
$\mathbf{1}_{\text{condition}}$	is 1 if the condition is true, 0 otherwise
\oplus	Concatenating two vectors
$\{t_i\}_{i=1}^N$	A sequence of N elements, such as tokens or vectors

Abbreviations

cf.	confer/conferatur (compare)
e.g.	exemplum gratia/conferatur (example)
et al.	et alia (and others)
etc.	et cetera (and so on)
i.e.	id est (that is)

ADE	Adverse Drug Event
ADR	Adverse Drug Reaction
BERT	Bidirectional Encoder Representations from Transformers
BIO	Beginning-Inside-Outside
EHR	Electronic Health Record
ELECTRA	Efficiently Learning an Encoder that Classifies Token Replacements Accurately
ELMo	Embeddings from Language Models
IE	Information Extraction
LSTM	Long Short-Term Memory
LM	Language Model
NER	Named Entity Recognition
NLP	Natural Language Processing
UMLS	Unified Medical Language System

CHAPTER 1

Introduction

The growth rate in the amount of biomedical documents—such as scholarly articles, clinical notes and health forum discussions—is staggering. Unlocking information trapped in these documents can enable researchers and practitioners to operate confidently in the information world. Biomedical Information Extraction (IE) system aims to automatically extract structured information—such as biomedical concepts, attributes, events, and their relations—from unstructured text. Within an IE system, the first step is called Biomedical Named Entity Recognition (NER), the task of recognising biomedical names.

In this chapter, we first use a real world application—recognising adverse drug events from social media—as an example, to illustrate how a NER model can be used to extract useful information (Section 1.1). Next, we describe a unified architecture for the most popular sequence tagging based NER models, dividing sequence taggers into two components: (1) a mapping function that maps each token to a feature vector, and, (2) a classifier that predicts a sequence of tags given the input sequence of feature vectors (Section 1.1.1). Then, in Section 1.2, we identify three open challenges of applying state-of-the-art sequence taggers, enhanced by pre-trained language representation models, to recognise biomedical names: (1) complex structures—overlapping and discontinuity—occur often in biomedical names; (2) the training of sequence taggers requires large training sets which are usually difficult to obtain in the biomedical domain; and, (3) there is a discrepancy between publicly available language representation models pre-trained on generic data and target biomedical data. To deal with these challenges, we explore different research directions and make the following contributions: we propose a transition-based model for discontinuous NER; we develop a

cost-effective approach that nominates the suitable pre-training data; and we design several data augmentation methods for NER (Section 1.3).

1.1 Recognising Adverse Drug Events from Social Media—A Motivating Application

An *Adverse Drug Reaction* (ADR) is an injury occurring after a drug is used at the recommended dosage, for recommended symptoms (Karimi et al., 2015b). Detecting ADRs as early as possible can potentially have a major impact, because ADRs are among the leading causes of death in many countries, and ADR-related costs have exceeded the cost of medications (WHO, 2020). Bonn (1998); Hadi et al. (2017); and Khalil and Huang (2020) estimate that ADRs account for more than 100,000 deaths per year in the United States, and 197,000 deaths annually in Europe. The situation in developing countries may be more severe. For example, Mouton et al. (2015) estimate that in South Africa ADRs contribute to the death of 2.9% of medical admissions, and 16% of deaths are ADR-related.

Different from controlled clinical trials which are mainly conducted *before* drugs are licensed for use, pharmacovigilance—the practice of monitoring the ADRs of pharmaceutical products—focuses on identifying previously unreported adverse reactions *after* the drugs are marketed. Establishing causality—whether the given adverse reaction is caused by the drug—is often done by domain experts. Causality assessment needs to investigate the statistical association in laboratory parameters and exclude other causes, such as alcohol, disease-related causes, other drugs and so on (Anderson and Borlak, 2011). Surveillance systems, both passive and active, play an important role in collecting potential *Adverse Drug Events* (ADEs). Note that, when causality between an adverse reaction and a drug is not known, it is referred to as an adverse drug event.

Passive surveillance of ADEs relies on spontaneous reporting systems which allow health professionals and patients to voluntarily report observed or suspected ADEs to regulatory

agencies. For example, the MedWatch system has been built by the Food and Drug Administration since the early 1990s (Piazza-Hepp and Kennedy, 1995). However, under-reporting is severe. Studies estimate that more than 90% of ADEs are not reported to these systems due to various obstacles, such as lack of suspicions, lack of information about reporting utility, lack of time, and difficulties in filling out forms (Vallano et al., 2005; Hazell and Shakir, 2006).

Active surveillance, in contrast, aims to discover ADEs automatically from multiple sources, including Electronic Health Records (EHRs), medical literature, search engine logs, and even social media. Since such information is often trapped in free text representation, IE systems can be used to extract information of interest. NER is usually employed at the very beginning of the IE system. The sentence, represented as a sequence of tokens, is taken as input of the NER model, and entity mentions, each of which is represented as a set of token positions, are outputted. In addition, one entity category, such as drug, disease, symptom, ADE and so on, is assigned to each entity mention.

A simple example. In this section, we describe a simple example of a post from a patient forum and explain how the NER model recognises biomedical names.

Given a sequence of tokens:

After two days of being on Cymbalta , I noticed an increase in flatulance¹
and the worst smelling gas I've ever smelled .

the NER model is supposed to recognise three entity mentions: '*Cymbalta*', as a drug mention, '*increase in flatulance*' and '*smelling gas*', as ADEs (Figure 1.1).

Sequence tagging based NER model. The state-of-the-art NER model is based on sequence tagging technique that assigns a tag to each token. The tag is usually composed of a position indicator and an entity category. The position indicator is used to represent the token's role in a mention. For example, in the BIO schema (Sang and Meulder, 2003), B stands for the Beginning of a mention, I for the Inside of a mention, and O for Outside a

¹The spelling error is from the original post.

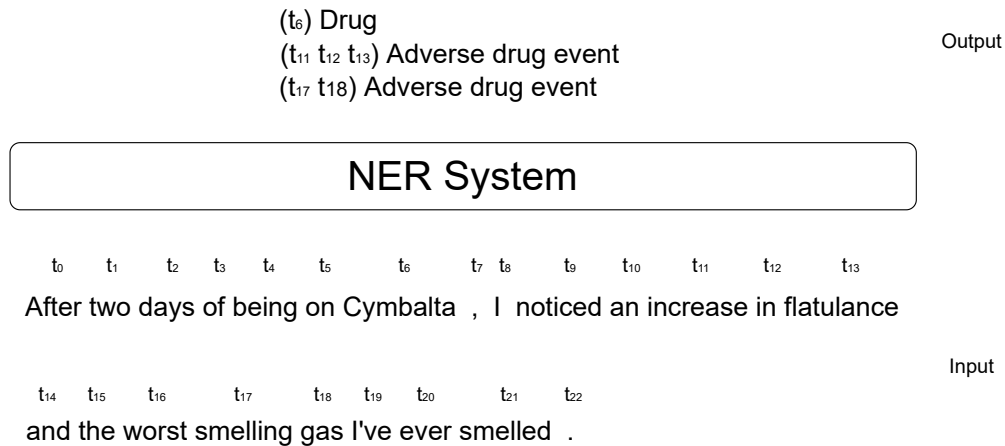


FIGURE 1.1. An example input sentence and the entity mentions which are supposed to be recognised by the NER model.

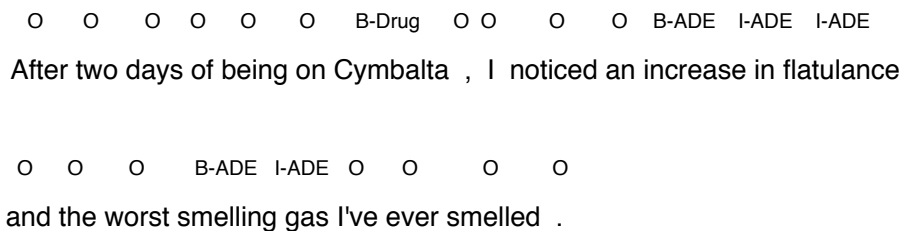


FIGURE 1.2. State-of-the-art NER model is based on sequence tagging technique that assigns a tag to each token. Token positions of mentions can be extracted from the output tag sequence.

mention. Figure 1.2 is an example of input sequence of tokens and the corresponding output sequence of tags. Taking the token ‘*smelling*’ as an example, its tag ‘B-ADE’ indicates that the token is the beginning token of an ADE mention.

Once the sequence of tags is outputted, token positions of mentions can be extracted from the tag sequence via finding all sub tag sequences starting with ‘B-∗’ tag, and including all succeeding ‘I-∗’ tags. Put another way, for each token whose tag starts with ‘B’, there is a mention starting at this token position, and ending before the next token position where the corresponding tag is ‘O’ or starts with ‘B’. Note that it is possible for the sequence tagger to predict an invalid sequence of tags, for example, a tag ‘B-ADE’ followed by a tag ‘I-Drug’. Therefore, post-processing steps, such as changing the tag’s position indicator ‘I’ to ‘B’ if its entity category is different from the preceding category, are usually employed before the tag sequence is decoded into mentions.

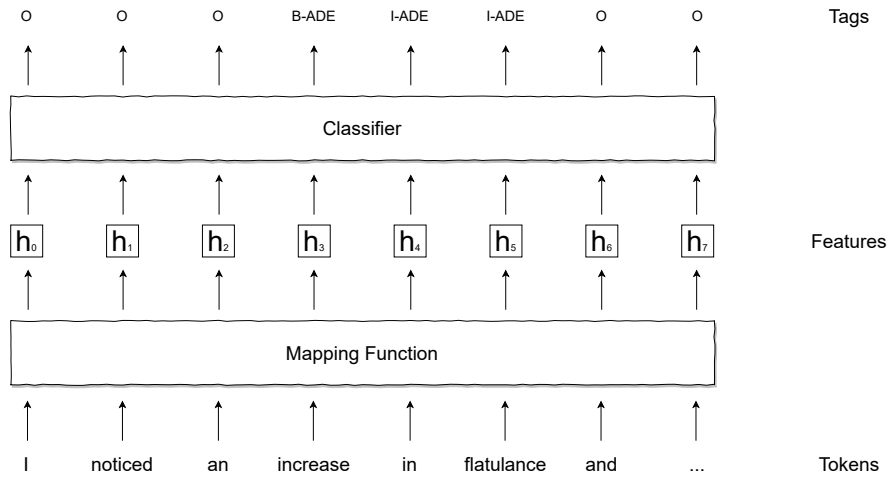


FIGURE 1.3. A unified architecture for sequence tagging models, consisting of a mapping function and a classifier.

1.1.1 A unified architecture for sequence taggers

In general, sequence taggers can be divided into two components (Figure 1.3):

Mapping function: it converts the input sequence of tokens into a sequence of features vectors, each of which represents the corresponding token-in-the-context; and,

Classifier: it predicts a sequence of tags given the input sequence of feature vectors.

The key to supervised machine learning based techniques is optimising the model so that they can fit the labelled training data. In other words, for each training instance consisting of the input sequence of tokens and the output sequence of tags, the tagger aims to predict a sequence of tags as close as possible to the grounding truth. The main advantage of recent deep learning based techniques over conventional feature based machine learning techniques is that the former optimises mapping function and the classifier jointly, whereas the mapping function in the latter is usually handcrafted and fixed during the model training stage.

Current state-of-the-art approaches (Huang et al., 2015; Lample et al., 2016; Ma and Hovy, 2016; Yang et al., 2018) for sequence tagging use the Bidirectional Long Short-Term Memory (BiLSTM) as the mapping function, and a subsequent linear-chain Conditional Random Field (CRF) as the classifier. FLAIR (Akbik et al., 2018) is a variant of BiLSTM-CRF sequence

tagger, which achieves the state-of-the-art performance in multiple sequence tagging datasets, including the CONLL 2003 English and German NER datasets. In this section, we detail its components in a top-down manner.

Linear-chain CRF. Given a sequence of feature vectors: $\{\mathbf{h}_i\}_{i=1}^n$, each of which representing a token-in-the-context, the simplest classifier can take each feature vector as input and makes the prediction independently. That is, for each feature vector at position i ,

$$\mathbf{o}_i = \text{softmax}(\mathbf{W}\mathbf{h}_i + \mathbf{b}). \quad (1.1)$$

This classifier ignores the relationship between neighbouring tags. For example, if the tag at a position is ‘*B-ADE*’ (beginning token of an ADE), it is impossible for the succeeding tag to be ‘*I-Drug*’ (inside token of a drug name), because the tag ‘*I-Drug*’ should always follow a ‘*B-Drug*’ (beginning token of a drug name) or another ‘*I-Drug*’.

CRF is a classifier that predicts the output sequence jointly, taking the dependency between neighbouring outputs into consideration. That is, it aims to predict a sequence of tags $\hat{\mathbf{O}} = \{\mathbf{o}_i\}_{i=1}^n$ which has the maximum probability over all possible tag sequences:

$$\hat{\mathbf{O}} = \arg \max_{\mathbf{O}} P(\mathbf{O} | \mathbf{H}), \quad (1.2)$$

where

$$P(\mathbf{O} | \mathbf{H}) \propto \prod_{i=1}^n \psi(\mathbf{o}_{i-1}, \mathbf{o}_i, \mathbf{h}_i) \quad (1.3)$$

and

$$\psi(\mathbf{o}_i, \mathbf{o}_j, \mathbf{h}) = \exp(\mathbf{W}\mathbf{h} + \mathbf{A}_{\mathbf{o}_i, \mathbf{o}_j}). \quad (1.4)$$

In Equation 1.4, $\mathbf{A}_{i,j}$ is the compatibility score of a transition from the tag i to tag j .

Token level BiLSTM layer. The LSTM variant of recurrent neural networks (Hochreiter and Schmidhuber, 1997; Graves et al., 2013) is widely used by recent work to create the contextual representation, due to its ability to flexibly encode long-term dependencies via a memory cell. In the LSTM architecture (Figure 1.4), the output state at each position (\mathbf{h}_i) is

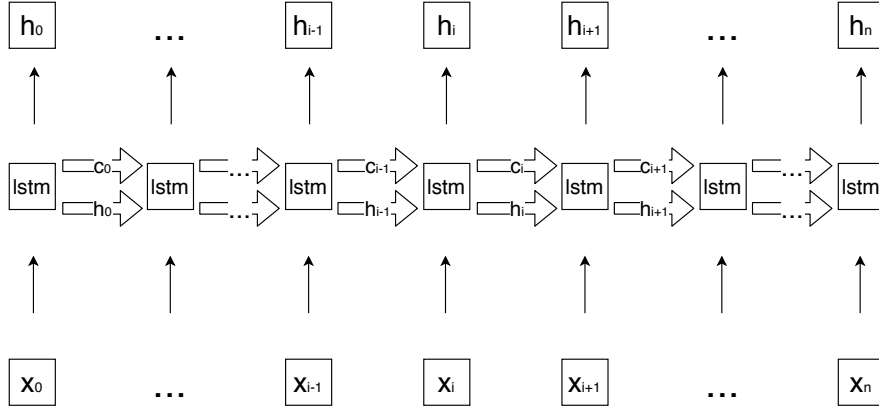


FIGURE 1.4. An LSTM computes the output state by taking the entire past (left context) of the input sequence into consideration.

computed by taking the input at the current position (x_i) as well as the hidden state and cell state from the previous position (h_{i-1} and c_{i-1} , respectively) into consideration:

$$h_i = f(t_i, h_{i-1}, c_{i-1}; \theta). \quad (1.5)$$

By computing recursively the hidden state, the entire past history—left context—of each position is incorporated. The term *bidirectional* indicates that there are two models—forward and backward models—used to capture both left and right contexts. The backward model works in the same way but in the reversed direction:

$$h_i = f(t_i, h_{i+1}, c_{i+1}; \theta). \quad (1.6)$$

In the following, we use the superscript f to define states relating to the forward model and b to the backward model. For example, h_i^f indicates the contextual representation of the i -th token from the output in Equation 1.5 and h_i^b the contextual representation of the i -th token obtained from output in Equation 1.6. A convention of employing BiLSTM is that the final contextual representation at each position h_i —the contextual representation of the i -th token—is usually extracted by concatenating the hidden states for each position from both forward and backward models:

$$h_i = \left[h_i^f \oplus h_i^b \right]. \quad (1.7)$$

	I	noticed	an	increase	in	flatulence	and	...
Token level indices	0	1	2	3	4	5	6	7
Character level start indices	0	2	10	13	22	25	37	41

TABLE 1.1. The sequence of tokens is treated as a sequence of characters.

Contextual string embeddings. FLAIR introduces a novel type of token embeddings based on character level encoder. The input token is first treated as a sequence of characters. Table 1.1 is an example sequence of tokens and the corresponding character level start indices.

Then the sequence of characters is taken as input of two—forward and backward—pre-trained character level BiLSTM models. The final contextual string embeddings for each token can be extracted by concatenating outputs from these two BiLSTM models. Specifically, for the forward model, the output hidden state after the last character in the token is used, and the output hidden state before the token’s first character from the backward model is used. Taking the token ‘*increase*’ in Table 1.1 as an example, the output state of the 12-th character from the backward model and the output state of the 21-st character from the forward model are concatenated as the contextual string embedding (Figure 1.5).

Finally, the stacking embeddings, a concatenation of contextual string embedding and pre-computed GLOVE embedding (Pennington et al., 2014), are used as the final token embedding and taken as input to the previous described token-level BiLSTM layer.

Pre-trained language representation models. In contrast to supervised machine learning that optimises model using labelled data only, semi-supervised learning aims to make use of both labelled data and unlabelled data. Pre-training language representation models on unlabelled data and then adapting pre-trained model to the downstream supervised task is one type of semi-supervised learning. It has demonstrated its effectiveness in NLP during the past decade (Mikolov et al., 2013b; Dai and Le, 2015; Howard and Ruder, 2018; Peters et al., 2018; Devlin et al., 2019). In this section, we briefly describe the design in FLAIR and more options are discussed in Section 2.3.1.

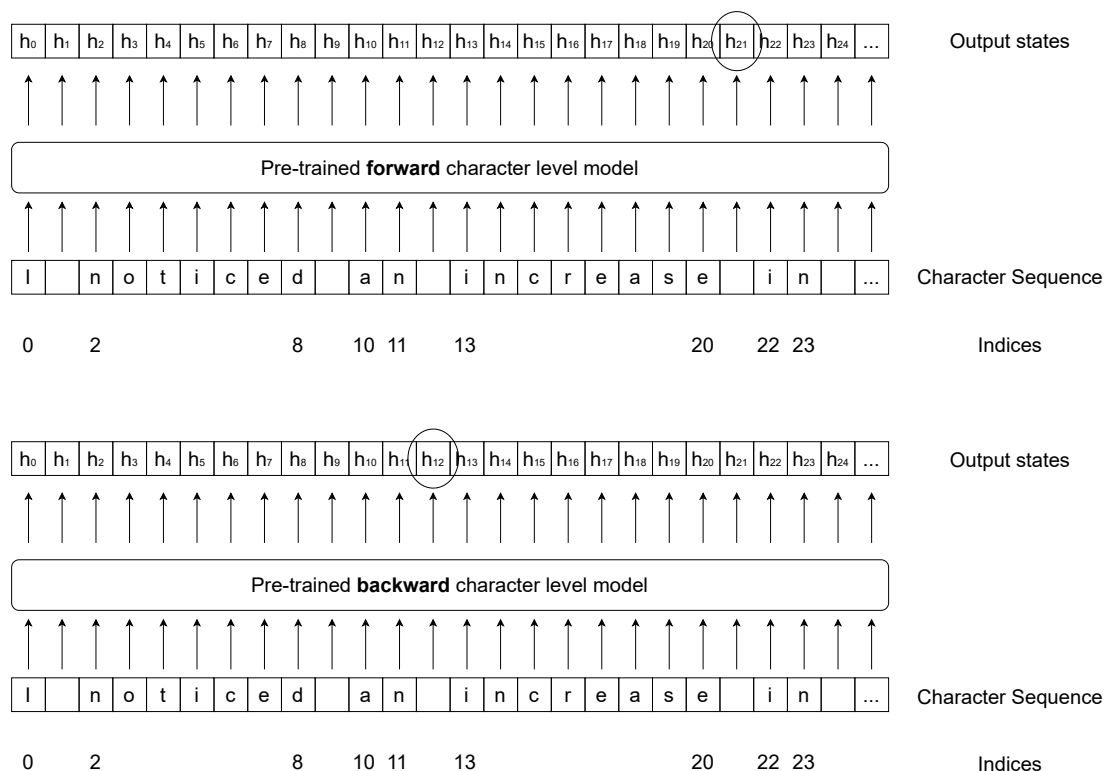


FIGURE 1.5. An forward model computes the output state by taking the entire past (left context) of the input sequence into consideration, whereas the backward model considers the entire right context. Both output hidden states are concatenated to form the final contextual string embedding and capture the information of the token itself as well as its surrounding tokens.

During the pre-training stage, Akbik et al. (2018) train two separate models—forward and backward—on the 1-billion word corpus (Chelba et al., 2013). The pre-training task is a standard character level language modelling task that predicts the next character given a sequence of characters. Taking the backward model illustrated in Figure 1.5 as an example, the output state of the 12-nd character is taken as input to a classifier to predict the next character, whose ground truth in this example is the character ‘n’. Once the pre-training finishes, the pre-trained models are frozen and used as part of the mapping function for downstream supervised task (Figure 1.6).

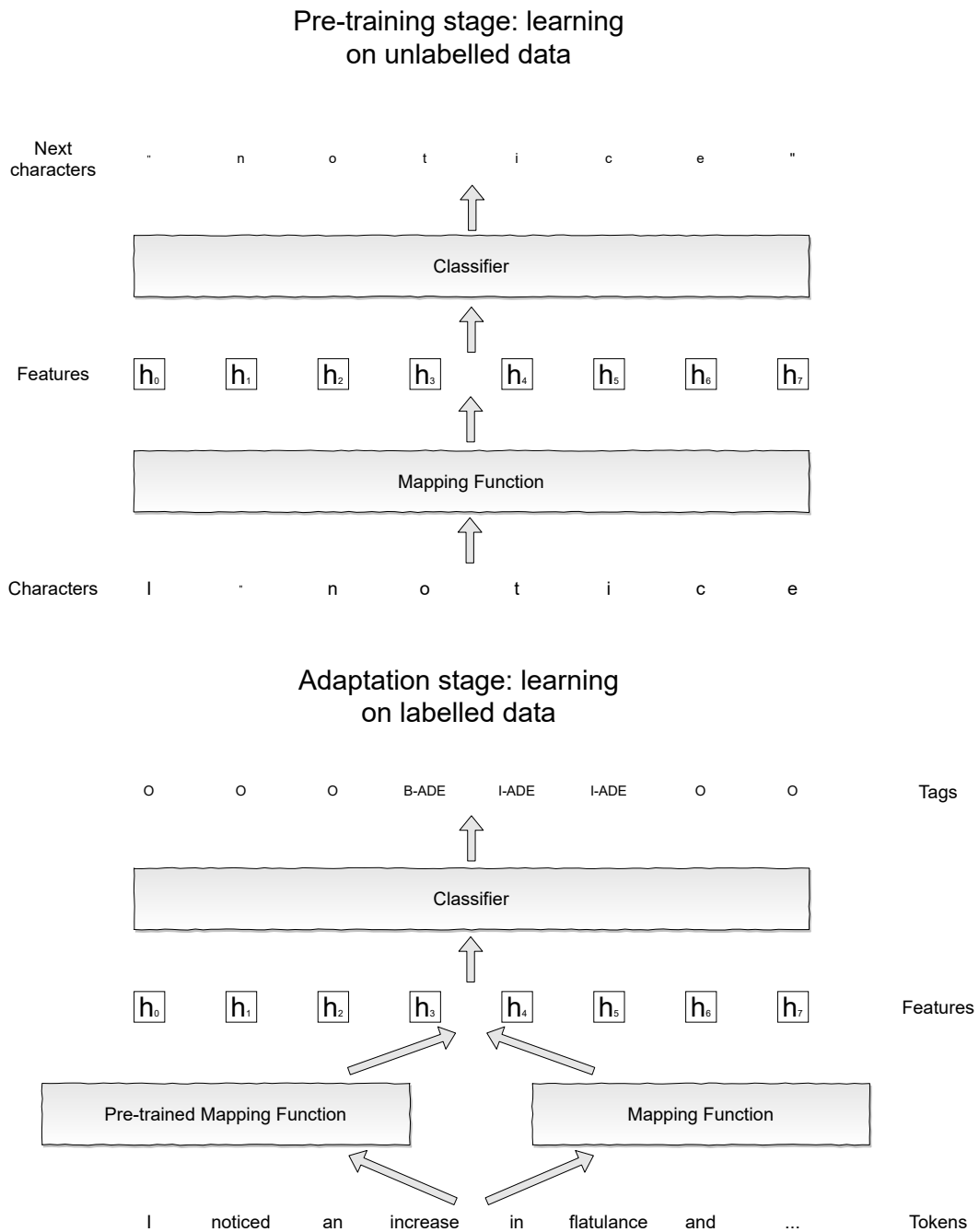


FIGURE 1.6. The semi-supervised learning approach used in FLAIR: pre-training two—forward and backward—character level language models, and using the pre-trained model as part of the mapping function in the downstream supervised tasks. For the sake of brevity, we show only the forward model and the backward model is omitted. The whitespace character is represented using " in this figure.

1.2 Key Open Challenges

Although FLAIR, enhanced by pre-trained language representation models, has achieved state-of-the-art performance in multiple NER datasets in the generic domain, we find there are several open challenges of applying FLAIR to recognise biomedical names.

1.2.1 Biomedical names are complex

Different from entity mentions in the generic domain, which are usually short spans of text, biomedical names may contain more complex inner structure. Considering the following sequence of tokens:

have much muscle pain and fatigue .

it contains two biomedical names, ‘*muscle pain*’ and ‘*muscle fatigue*’, that share the token ‘*muscle*’. In this case, we call them *overlapping* biomedical names. In addition, ‘*muscle fatigue*’ is a *discontinuous* mention, consisting of two components that are separated from each other.

The main motivation for recognising these biomedical names with complex inner structure is that they usually represent compositional concepts that differ from concepts represented by individual components. Specifically, each of these two names in the example sentence—‘*muscle pain*’ and ‘*muscle fatigue*’—describes a disorder which has its own CUI (Concept Unique Identifier) in UMLS (Unified Medical Language System), whereas ‘*muscle*’, ‘*pain*’, and ‘*fatigue*’ also have their own CUIs. In downstream applications, such as pharmacovigilance, extracting these compositional concepts, such as symptoms or ADEs, is often more useful than extracting individual components which may refer to body locations or general feelings.

1.2.2 Labelling data is difficult

A well known limitation of deep neural models is that training these models usually requires large amount of labelled data (LeCun et al., 2015). In other words, the advantages of deep

learning may diminish when working with small training sets. For example, Shen et al. (2018) observe that a deep neural model outperforms the best shallow model by absolute F_1 score of 2.2, when a large NER training set—ONTONOTES 5.0, containing more than 1 million tokens—is available. In contrast, this advantage becomes only 0.4, when training on a comparatively small training set (CONLL 2003, containing around 0.2 million tokens).

Labelling large amount of generic NER data is time-consuming, because the annotation needs to be done at the token level. Labelling large amount of biomedical NER data is even more difficult due to the following reasons:

- The previously described complex structure increases the difficulty of annotation. Standard NER annotation is usually done at the token level: annotators need to scrutinise every token to decide whether it is part of *one* entity mention. However, due to the complex structure—overlapping and discontinuity—in biomedical names, one token may belong to multiple biomedical names, and tokens that are far away from each other may form one biomedical name. Exhaustive enumeration of possible names, including discontinuous and overlapping ones, is exponential to sentence length.
- Domain knowledge is required to annotate biomedical NER datasets. Different from the task of annotating generic entity mentions, such as person names or locations, with which ordinary people are familiar, recognising biomedical names, such as biological substances or disorders, requires the annotators to have at least basic domain knowledge.

Worse still, the same entity category may have subtle meanings in different biomedical applications. This may even require annotators to have expert level knowledge in a specific application. For example, *family history extraction* is a task that focuses on the detection of family history related disorders. Therefore, it also pays attention to some behaviour patterns which may be caused by genetic factors, and these behaviour patterns are usually overlooked by popular disorder recognition tasks (Rybinski et al., 2021). In other words, a behaviour pattern is usually not defined as a disorder of interest in most of biomedical applications, but

needs to be labelled as disorder in family history extraction application, once the behaviour—such as a pattern of alcohol use—may put people health at risk, and it may be influenced by genetic factors.

- Some annotation tasks in the biomedical domain may cause negative impacts on annotators. For example, annotators may feel uncomfortable after continuing annotating online posts about adverse drug events for a long time. These posts are written by patients, containing complains about their sufferings after drug usage. Proper protective arrangements need to be made to protect the annotators, and they usually lead to longer annotation task duration.
- The last, but not the least, reason relates to the cost-benefit analysis widely used in project management activities. That is, a project for building biomedical applications usually starts from defining target performance specifications, and then estimates the cost of achieving the target performance. Labelling training data is often the most expensive part of the project, and, unfortunately, we do not have practical methods to estimate how much training data is required to achieve the target performance (Johnson et al., 2018). So a more practical strategy is that domain experts usually first annotate a small set of training data, on which NLP practitioners need to build pilot models. After the persuasive results are obtained using limited amount of training data, domain experts and project managers are more likely to commit more resources to create more labelled training data.

1.2.3 Unlabelled biomedical data are limited

The main strengths of FLAIR come from the use of stacking embeddings, that consist of two types of embeddings: pre-trained GLOVE embedding (Pennington et al., 2014) and contextual string embeddings based on pre-trained language models. Akbik et al. (2018) show that the use of pre-trained GLOVE embedding increases average F_1 score by 1.1, and the use of contextual string embeddings brings even larger improvements, around 4.5 absolute F_1 score on the English NER dataset. However, both of these two types of embeddings are pre-trained on generic data. For example, GLOVE embeddings are pre-trained on the English Wikipedia

	CoNLL 2003	SHARE/CLEF 2013
GLOVE (Vocabulary coverage)	87.6 % 18,415 / 21,089	37.2 % 5,282 / 14,172
String embeddings (Perplexity)	7.746	29.839

TABLE 1.2. Discrepancy between the pre-trained model used in FLAIR and the target datasets: SHARE/CLEF 2013 (clinical notes) and CoNLL 2003 (news stories).

and Gigaword dataset (archive of news stories). These generic data usually have very different characteristics from the biomedical data.

Pre-trained models used in FLAIR usually have sub-optimal performances on biomedical datasets, such as SHARE/CLEF 2013, which is sourced from clinical notes.

We measure this discrepancy between the pre-trained model and the target data using two measures: vocabulary coverage and perplexity. Vocabulary coverage indicates the ratio of target data’s vocabulary existing in the pre-trained model. For example, there is only 37.2 % of SHARE/CLEF 2013’s vocabulary covered by GLOVE (Table 1.2). Perplexity is a way of evaluating the language model, which is used to generate contextual string embeddings in FLAIR. The pre-trained character level language models achieve higher perplexity—a measurement of how well a language model predicts a test sentence—on SHARE/CLEF 2013. Note that high perplexity indicates the language model is bad at predicting the test sentence, assigning low probability. The result suggests that there is a higher discrepancy between pre-trained models and the SHARE/CLEF 2013 than CoNLL 2003, which is sourced from news stories.

Unfortunately, the access to unlabelled data in the biomedical domain can be restricted due to privacy and regulatory reasons. Documents with privacy sensitive contents, such as electronic health records, are usually available only after applying anonymisation operations. For example, the Health Insurance Portability and Accountability Act (HIPAA) of the United States defines that 18 types of Protected Health Information (PHI), such as patient names, ages, phone numbers etc., need to be removed from the documents before they can be shared

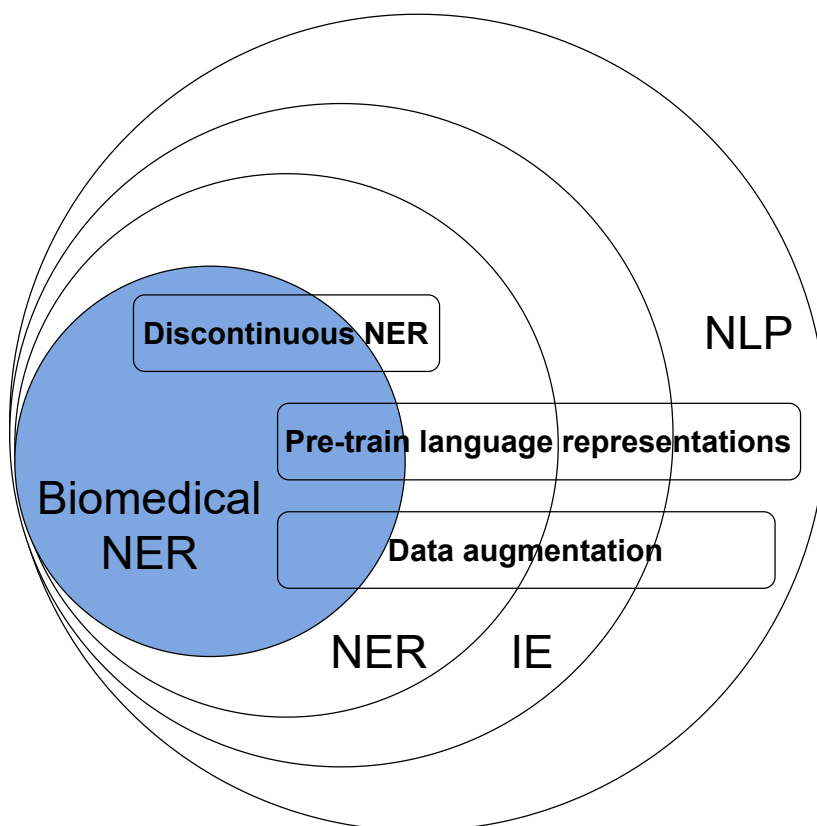


FIGURE 1.7. We focus on Biomedical NER, and explore the following three research directions: recognising discontinuous entity mentions; pre-training domain-specific language representation models; and enhancing the effectiveness of NER models using data augmentation.

with third parties. Selecting proper pre-training data which are large enough and also similar to target task data is a non-trivial problem.

1.3 About the Thesis

To deal with these open challenges, we explore the corresponding research directions, aiming to improve the Biomedical NER. Figure 1.7 is a high-level overview of concepts we cover in this thesis. Although we focus on Biomedical NER in this thesis, some of these contributions, including proposed discontinuous NER models and new discoveries regarding the selection of pre-training data, can be potentially applied to other NLP tasks in other domains. Also it is worthy noting that we consider English text only.

1.3.1 Publications

The work in the thesis primarily relates to the following peer-reviewed articles (sorted by publication date):

- (1) **Xiang Dai**, Sarvnaz Karimi, and Cecile Paris. 2017. Medication and adverse event extraction from noisy text. In Proceedings of the Australasian Language Technology Association Workshop, pages 79–87, Brisbane, Australia. (Chapter 1)
- (2) **Xiang Dai**. 2018. Recognizing complex entity mentions: A review and future directions. Proceedings of ACL 2018, Student Research Workshop, pages 37–44, Melbourne, Australia. (Section 2.4)
- (3) **Xiang Dai**, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2019. Using similarity measures to select pretraining data for NER. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1460–1470, Minneapolis, Minnesota. (Chapter 4)
- (4) **Xiang Dai**, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An effective transition-based model for discontinuous NER. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5860–5870, Online. (Chapter 5)
- (5) **Xiang Dai**, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. Cost-effective Selection of Pretraining Data: A Case Study of Pretraining BERT on Social Media. In Findings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online. (Chapter 4)
- (6) **Xiang Dai**, Heike Adel. 2020. An Analysis of Simple Data Augmentation for Named Entity Recognition. In Proceedings of the 28th International Conference on Computational Linguistics, Online. (Chapter 3)

The following articles are related, but will not be extensively discussed in this thesis:

- (7) Nicky Ringland, **Xiang Dai**, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R. Curran. 2019. NNE: A dataset for nested named entity recognition in English newswire. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5176–5181, Florence, Italy.
- (8) Aditya Joshi, **Xiang Dai**, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2018. Shot or not: Comparison of NLP approaches for vaccination behaviour detection. In Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task, pages 43–47, Brussels, Belgium.
- (9) Lukas Lange, **Xiang Dai**, Heike Adel, Jannik Strötgen. 2020. NLNDE at CANTEM-IST: Neural Sequence Labeling and Parsing Approaches for Clinical Concept Extraction. In Iberian Languages Evaluation Forum (IberLEF 2020), Online.

1.3.2 Definition and clarification of used terms

The usage of technical terminologies in the literature is usually confusing and inconsistent, especially when researchers from different communities use the same term to refer to different concepts, or when some conventions are only shared by a small group of people. For the sake of brevity, we define and clarify some frequently used terms in this thesis.

Token: An individual occurrence of a linguistic unit in text. We use a token to refer to an individual word unless specified otherwise. If a word is further split into several pieces, we use sub-tokens to refer to these pieces.

Span: A consecutive sequence of tokens, or an individual token.

Biomedical concept: Conceptual objects, events, and procedures in the biomedical ontologies. We use entity and biomedical concept interchangeably.

Biomedical name: An instance where a biomedical concept is referenced to in text. We use mention and biomedical name interchangeably. Following (McDonald et al., 2005), we denote the mention by the set of token positions that belong to the mention. Therefore, a mention may consist of several spans and mentions may overlap.

Embedding: A mapping function that converts a token into a dense vector. It can also be considered as a look-up dictionary, where the token is the key and the vector is the value.

Encoder: A mapping function that converts a sequence of tokens into a sequence of dense vectors:

$$[\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N] = \text{Text encoder}([t_1, t_2, \dots, t_N]),$$

where N is the sequence length.

Usually, the encoder is a trainable neural network, and the output vectors are contextualised in the sense that they reflect both the corresponding token and its contexts.

Attention: A mechanism that is widely used in sequence models to allow the current state ‘attending’ to context states to obtain a context vector. The resulting context vector is usually be used together with the current state for the downstream layers. Given the current state \mathbf{h}_i and a sequence of context states $\mathbf{H} = \{\mathbf{h}_j\}_{j=1}^N$, we can calculate the context vector \mathbf{c}_i as the average of context states weighted with attention scores:

$$\mathbf{c}_i = \sum_{j=1}^N a_{ij} \mathbf{h}_j, \quad (1.8)$$

where a_{ij} is the j -th element in the attention vector a_i :

$$a_i = \text{softmax}(f(\mathbf{h}_i, \mathbf{H})). \quad (1.9)$$

We use the function f proposed by Luong et al. (2015), unless specified otherwise:

$$f(\mathbf{h}_i, \mathbf{H}) = \mathbf{h}_i^\top \cdot \mathbf{W} \cdot \mathbf{H}, \quad (1.10)$$

where \mathbf{W} is a trainable weight matrix.

Language representation model: Similar to text encoder, a language representation model can generate a contextual vector representation for each input token. The subtle difference between a text encoder and a language representation model is that the former focuses more on generating task-oriented representations, and the latter

emphasises the general semantic and syntactic representations. For example, a text encoder in an NER model may assign similar vectors to words belonging to the same entity category, even though their semantic meaning are dissimilar.

Domain-specific vs. Generic domain: The term ‘domain’ is loosely used in both machine learning and NLP communities (Ramponi and Plank, 2020). In this thesis, we do not attempt to define what constitutes a domain but assume the domain exists in a covert way.

We use domain-specific model to indicate that the model focuses on a specific domain. For example, we call the language representation model trained on biomedical corpora a domain-specific model. In contrast, we use generic model to refer to a model which is supposed to capture any kind of knowledge. For example, we call the language representation model trained on Common Crawl corpus a generic domain model.

Supervised vs. Unsupervised: We use supervised learning to indicate that human annotators are required to label the training set. In contrast, if training does not require labels annotated by human annotators, then we call it unsupervised learning, or self-supervised learning. For example, we call pre-training language models unsupervised learning. The task is to predict the next token, given a sequence of tokens, and this task does not require human annotated labels.

1.4 Summary

In this chapter, we first illustrate how an NER model can be used to extract useful information, improving applications in the biomedical domain. Next, we describe a state-of-the-art NER model, FLAIR, which is based on sequence tagging techniques. Then, we identify three open challenges of applying current techniques to recognise biomedical names.

In the following chapters, we organise our related work and content chapters into three groups, each of which focuses on solving one particular challenge: Section 2.2 and Chapter 3 focusing on the lack of labelled data problem; Section 2.3 and Chapter 4 on selecting suitable

pre-training data given the downstream task; and Section 2.4 and Chapter 5 on recognising biomedical names containing complex inner structure.

Literature Review

Named Entity Recognition (NER) techniques have developed gradually from dictionary based and rule based to machine learning based approach during the last several decades. Motivated by previously discussed challenges of training supervised models—e.g., labelling biomedical NER dataset can often be expensive and time-consuming—we provide an overview of promising approaches to overcome the lack of training data problem, with a special focus on data augmentation (Section 2.2.3) as well as transfer learning (Section 2.3). Additionally, complex structures—overlapping and discontinuity—are common in biomedical names. We review the existing methods for complex entity recognition, and group these methods into token-level, span-level and sentence-level approaches (Section 2.4).

2.1 A Brief History of NER

Information Extraction (IE) is an important Natural Language Processing (NLP) task that aims to automatically extract structured information from unstructured text. It has been widely used in many applications. For example, a successful email system can identify messages that contain event information, extract the attributes of the event (i.e., time, location, and participants), and insert the extracted event to the calendar (Laclavík et al., 2012). In the biomedical domain, IE has been widely used to extract biomedical concepts, attributes, events, and their relations from scholarly articles, clinical notes, and social media data (Sarawagi, 2008; Wang et al., 2018b).

One of the common practices in IE is to separate processing into several stages, among which NER is typically employed as the first step (Hobbs, 2002). On the one hand, NER

requires a deeper analysis than key word searches, because the semantics of entity mentions are influenced by their contexts. For example, ‘*Washington*’ may refer to a person, a city, a state, or an organisation, depending on the contexts. On the other hand, NER does not seek to fully understand every aspect of the text, such as the writer’s communicative intent (Bender and Koller, 2020). Therefore, it focuses only on relevant words and ignores the rest. Because of its location at a midpoint on this spectrum, NER is a fundamental task, and it has received considerable attention in the last several decades.

Grishman and Sundheim (1996) use the term *Named Entity*, referring to possible persons, organisations, and locations mentioned in text, and they aim to recognise structured information of company activities and defence related activities from newspaper articles. Florian et al. (2004) extend the task to recognise mentions of textual references to conceptual objects, which can be either named (e.g. ‘*George Washington*’), nominal (e.g. ‘*The president*’) or pronominal (e.g. ‘*He*’). The entity categories studied in the generic domain are mainly person, organisation, and location.

Biomedical NER, focusing on identifying and classifying biomedical names, whose surface forms can represent biomedical concepts, has its unique characteristics comparing to NER in the generic domain. Early stage efforts, e.g., GENIA project (Collier et al., 1999; Kim et al., 2003) and BioCreAtIvE (Critical Assessment for Information Extraction in Biology) challenges (Hirschman et al., 2005), focus on automatically extracting genome information from biochemical papers written by domain specialists. i2b2 (Informatics for Integrating Biology & the Bedside) and n2c2 (National NLP Clinical Challenges) projects (Kohane et al., 2006; Brownstein et al., 2010) start to bring Electronic Health Records (EHRs) to researchers’ attention, by releasing publicly available de-identified clinical notes. Additionally, the value of informal sources, such as user generated text on the web and search engine logs, have also been recognised by researchers. They start to use these data for mining health related information, such as predicting epidemic events (Joshi et al., 2019), and monitoring adverse drug events (Sarker et al., 2015).

In this section, we provide a brief overview of the development of NER techniques. Instead of exhaustively surveying different approaches and discussing design variants, we describe

representative work and focus on identifying what are the strengths and limitations of different approaches. For more detailed surveys of NER techniques in both generic and biomedical domains, we refer the reader to (Nadeau and Sekine, 2007; Campos et al., 2012; Yang et al., 2018; Yadav and Bethard, 2019; Li et al., 2020).

2.1.1 Dictionary based approach

Mikheev et al. (1999) build a minimal NER system equipped with dictionaries, also known as gazetteers or name lists. They collect person names, organisation names and location names from the MUC-7 training data, as well as several external resources, including the CIA World Fact Book, financial web sites, etc. Despite its simplicity, evaluation results on MUC-7 test set show that pure list lookup—finding occurrences of exact matches with items from dictionaries—performs reasonably well for locations (precision of 0.90 and recall of 0.86), but not for the organisation and person categories (recall of lower than 0.50, precision of around 0.80).

One serious limitation of this approach is that it cannot recognise unseen entity mentions, i.e., entities not in the dictionaries. In addition, maintaining large dictionaries requires great efforts. For example, there are around 1.5 million unique family names, just in the United States. The dictionary of company names, if at all available, would be much larger and out of date quickly, because new companies emerge all the time.

Naming variation is another issue that needs to be overcome. For example, the organisation dictionary might contain '*University of Sydney*', but this organisation may also be referred to as '*Sydney Uni*'. In the biomedical domain, this problem is even more severe. For example, the drug '*Acetylcysteine*', usually used for cough and other lung conditions, is also known as '*Acetyl Cysteine*', '*Cysteine Hydrochloride*', '*Cystine*', '*N-acetyl cysteine*', '*N-acetylcysteine*', '*N-acetyl-L-cysteine*', '*N-Acétyl-L-Cystéine*', etc.

Finally, ambiguity may be caused by the overlapping between dictionaries belonging to different entity categories. For example, '*J. P. Morgan*' could belong to both the person name dictionary and the organisation name dictionary. Ambiguity can also be caused by the usage

of abbreviations and acronyms. For example, ‘*CRF*’ may refer to ‘*Conditional Random Field*’ in the context of natural language processing. However, the possible number of meanings of the term ‘*CRF*’ in the context of biomedical is much larger, including ‘*Cardiorespiratory fitness*’—relating to heart health, ‘*Clinical risk factors*’, ‘*Controlled Rate Freezer*’—a medical equipment, ‘*Chronic renal failure*’—a type of kidney disease, etc. Note that clinical notes are usually written by practitioners under time pressure. So abbreviations and acronyms are used frequently. All of these limitations make the dictionary based approach more difficult to be widely employed.

2.1.2 Rule based approach

To overcome the previously mentioned limitations of dictionary based approaches, efforts were made to handcraft a set of rules to alleviate the reliance on the completeness of dictionaries. Rules can be created to expand the dictionaries to identify previously unseen mentions. For example, MetaMap (Aronson, 2001; Aronson and Lang, 2010) makes use of external knowledge sources of biomedical terms—the SPECIALIST lexicon, and it employs complex rules to identify all possible mention variants of an entity, including acronyms, abbreviations, synonyms, or derivational variants. Table 2.1 is an example that illustrates how expansion rules are used to generate variants given a word. Expansion rules include ‘i’ (inflection), ‘p’ (spelling variant), ‘a’ (acronym/abbreviation), ‘e’ (expansion of acronym/abbreviation), ‘s’ (synonym) and ‘d’ (derivational variant). For example, the expansion rule of variant ‘ophthalmia’—‘ssd’—indicates that it is a derivational variant of a synonym (‘*ophthalmic*’) of a synonym (‘*eye*’) of ‘*ocular*’.

Rules can also be triggered by characteristic attributes of known entity mentions, including their spellings and the contexts in which they appear. For example, a spelling rule can be a simple look up for the string, such as *any string containing ‘Mr.’ is a person*; or a spelling pattern, such as *any all capitalised string is an organisation (e.g., ‘IBM’)*. A contextual rule gets clues from surrounding words and their syntactic relationships, such as *any proper name modified by an appositive whose head is ‘president’ is a person (e.g., ‘Maury Cooper’ in the context of ‘... says Maury Cooper, a vice president at ...’)*.

Origin	Variant			POS	Expansion Rule
ocular				adj	–
	eye			noun	s
		eyes		noun	si
		optic		adj	ss
		ophthalmic		adj	ss
			ophthalmia	noun	ssd
	oculus			noun	d
		oculi		noun	di

TABLE 2.1. The variants of word ‘ocular’ and the corresponding rules to generate them. The indentation reflects the hierarchical structure of these variants according to the history of how they are generated.

One advantage of a rule based approach is that rules can be derived using unlabelled text only, which are much easier to obtain. For example, Collins and Singer (1999) build a named entity classifier using 90,000 unlabelled examples. They start from 7 seed rules (*‘New York’, ‘California’ and ‘U.S.’ are locations; any name containing ‘Mr.’ is a person; any name containing ‘Incorporated’ is an organisation; ‘I.B.M.’ and ‘Microsoft’ are organisations*), which is the only supervision in their approach. The classifier, automatically inducing new spelling rules and contextual rules, finally achieves over 91% accuracy when evaluated on a test set of 1,000 manually labelled instances.

However, applying these rules is challenging, when the number of rules become large. That is, it is difficult to prioritise one particular rule over others, especially when some of these rules may conflict with each other. For example, the following three rules may be used to represent the same example: *any all capitalised string is an organisation (e.g., ‘IBM’)*; *any string which is all capitalised or full periods, and contains at least one period is a location (e.g., ‘N.Y.’)* and *any string has an appositive modifier whose head is a singular noun (‘player’) is a person (e.g., ‘L.J., the greatest basketball player’)*. Iterating over all possible applicable rules and arranging them in order of importance, even if at all possible, will cause heavy computations.

Another difficulty of applying these rule based systems is that they usually rely on other NLP tools, such as a syntactic parser. For example, Zhang and Elhadad (2013) use a noun phrase chunker to first identify candidate entity mentions; and context rules used by Collins and

Singer (1999) involve finding the head word of the appositive modifier for the entity mention. Building syntactic analysis tools itself is a challenge task, especially for syntactically noisy text, such as clinical notes and social media data.

2.1.3 Statistical machine learning based approach

Statistical machine learning approaches replace ‘hard’ rules with ‘soft’ features and estimate the importance (weights) of features using labelled training data. Tokens are typically represented by vectors, each of which can consist of boolean, numeric and nominal values, representing each token-in-the-context. For example, a boolean value can be used to indicate whether the token is capitalised, and a nominal attribute can be used to represent the stem of the token. The feature creating function, mapping from a token to a sparse vector, is called a feature template. It controls the length of the token vector and the meaning of each element in the vector. Once the feature template is fixed, feature vectors—created via the same mapping function—can be taken as input of any supervised classifier, including Decision Tree, Maximum Entropy Models, Support Vector Machines, Hidden Markov Models and Conditional Random Fields.

Despite the successful applications of machine learning based NER, its main shortcoming is the requirement of sophisticated feature templates. These features should be informative and generalisable for unseen data. This is challenging because such high quality feature engineering requires expert domain knowledge and is usually tailored to specific entity categories or text types. Learning from these features may also suffer from the sparsity problem. For example, if a stem appears only one time in the training data, it is impossible to estimate its importance—the weight associated with the feature—from such a rare observation.

To alleviate the burden of manually building feature templates, deep learning models enable automated feature extraction. Distributed representations are usually employed to solve the sparsity issue. In other words, the mapping function is a neural network. It takes a token as input, and it outputs a dense vector instead of a sparse vector. Feature vectors—created via

Evaluation Dataset	F_1	Task Description
NCBI-DISEASE	87.5	Recognise disease names in biomedical publications
I2B2-2010	43.6	Recognise disease names (labelled as <i>problem</i>) in clinical notes
N2C2-2019	60.2	Recognise genetic disease names (labelled as <i>observation</i>) in clinical notes

TABLE 2.2. Decline in effectiveness of a model trained on NCBI-DISEASE (Doğan et al., 2014), when evaluated on other datasets: I2B2-2010 (Uzuner et al., 2011), and N2C2-2019 (n2c2, 2019). The mention-level F_1 score is reported.

the neural network—can be combined with almost any previous mentioned classifier, except for those which are better at sparse input vectors, such as Decision Tree.

2.2 NLP for Low Resource Scenarios

Although supervised neural models have achieved state-of-the-art performance on numerous benchmark NER datasets in the generic domain, due to the availability of large amount of labelled training data (*high resource*), they does not cover all applications. On one hand, the trained model usually does not generalise well across different types of text, let alone to recognise mentions belonging to new entity categories (Table 2.2). On the other hand, re-annotation for a new task, domain or language requires considerable effort.

In this section, we describe several approaches—except transfer learning which will be detailed in Section 2.3—to deal with the lack of labelled training data problem (*low resource*). We describe data augmentation approaches in details, because our methods described in Chapter 3 are built on top of such related work.

2.2.1 Distant supervision

Instead of manually labelling data, one research direction—often called *distant* or *weak* supervision (Hoffmann et al., 2011)—aims to automatically create training data by exploring existing knowledge base (e.g., Wikipedia, MeSH¹, CTD²) or heuristic rules.

Nothman et al. (2008) transform Wikipedia into named entity annotations by (1) classifying Wiki articles into common entity categories; (2) finding all possible inter-article links; and (3) assigning the entity category of the target page to the anchor text. Because the authors of Wikipedia are dictated to link only the first mention of an entity in each article, Nothman et al. use several heuristic rules to infer additional links from shorter referential forms. For example, the first or last word of a person name found later in the article may also refer to the same person. In addition, heuristic rules are employed to adjust link boundaries. For example, linked text may contain the possessive 's at the end of a name, and it should be removed from the entity name.

Safranchik et al. (2020) describe a framework, which takes unlabelled data and a set of rules as input, for creating labelled training data. Rules, which are implemented as functions, can take unlabelled data as input and output heuristic information about tags. For example, a simple rule can be '*tagging any token that appear in a dictionary of known entity category as I-*, and all other tokens as ABS, indicating that the rule abstains from assigning a tag*'. Taking the sentence in Figure 1.2 as an example, this rule—combined with a drug dictionary and an adverse drug event dictionary—may create the sequence of tags in Figure 2.1, if there are some tokens appearing in these dictionaries.

Note that it is possible that different rules output conflicting tags, if one token appears in multiple dictionaries. Also, it is possible that the identified span is incomplete. For example, '*increase in flatulance*' should be identified as an adverse drug event, but because '*increase in*' may not appear in the adverse drug event dictionary, these two tokens are labelled as '*ABS*' (the rule abstains from assigning a tag). To reconcile the conflicting and incomplete

¹Medical Subject Headings: <https://www.nlm.nih.gov/mesh/meshhome.html>. Accessed date: 22nd May 2021

²Comparative Toxicogenomics Database: <http://ctdbase.org/downloads/>. Accessed date: 22nd May 2021

Output 2	ABS	ABS	ABS	ABS	ABS	ABS	I-Drug	ABS	ABS	ABS	ABS	ABS
Output 1	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS
Input	After two days of being on Cymbalta , I noticed an increase											
Output 2	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS
Output 1	ABS	I-ADE	ABS	ABS	ABS	I-ADE	I-ADE	ABS	ABS	ABS	ABS	ABS
Input	in flatulence and the worst smelling gas I've even smelled .											

FIGURE 2.1. Example sequence of tags generated by a rule and two domain-specific dictionaries. For example, ‘*Cymbalta*’ is assigned the tag ‘*I-Drug*’ because it appears in the drug dictionary.

Evaluation Dataset	Precision	Recall	F_1
CONLL DUTCH (Sang, 2002)	32.4	21.1	25.5
CONLL SPANISH (Sang, 2002)	51.0	24.7	33.3
CONLL ENGLISH (Sang and Meulder, 2003)	39.9	30.1	34.3
CONLL GERMAN (Sang and Meulder, 2003)	23.2	9.2	13.2
ESTONIAN (Tkachenko et al., 2013)	59.7	49.3	54.0

TABLE 2.3. Evaluation results, as reported by Lange et al. (2019), of automatically annotated labels against manual annotations.

information, Safranchik et al. (2020) introduce a set of *linking rules* that decide whether adjacent tokens should be grouped into one span, and which tag is used for the span. For example, ‘*increase in flatulence*’ can be grouped into a span and share the tag assigned to ‘*flatulence*’. These linking rules are usually implemented based on automatic phrase mining, or with the help of language models that predicts which words may co-occur.

Although these described automatic labelling methods provide a cheap way to obtain a large amount of labelled training data, the obtained labelled data are usually *noisy*. Automatically annotated labels usually contain more errors than the manual annotations (Table 2.3), which are in contrast called *clean data*. Liang et al. (2020) point out that there is a trade-off between recall and precision using automatic labelling. That is, setting strict rules can generate high precision labels, but may not generalise well and thus have low recall. In contrast, relaxed rules can increase the coverage of annotation, leading to high recall and low precision.

Learning in the presence of noisy labels. Training a supervised model on noisy labels can sometimes result in negative results. Fang and Cohn (2016); Hedderich and Klakow (2018) show that training on the combination of noisy training data and a small amount of clean training data performs worse than training on clean data only. Therefore, efforts are made to solve the noisy labelled data problem (Han et al., 2018; Liang et al., 2020).

One popular approach of training with noisy labels is to model the true label as a latent variable and learn a noisy model that relate the true and noisy labels (Hedderich and Klakow, 2018; Lange et al., 2019). We use $P(y|x)$ to represent the probability distribution of a small set of clean instances $(x, y) \in \mathbb{C}$, and use $P(\tilde{y}|x)$ to represent the distribution of a large set of noisy instances $(x, \tilde{y}) \in \mathbb{N}$. Then, the noisy distribution can be calculated using:

$$P(\tilde{y} = j|x) = \sum_{i=1}^k P(\tilde{y} = j|y = i)P(y = i|x). \quad (2.1)$$

To estimate the relationship between true and noisy labels, i.e., $P(\tilde{y} = j|y = i)$, Hedderich and Klakow (2018) first apply the same auto-labelling operations on clean data \mathbb{C} to obtain pairs of clean y and corresponding noisy label \tilde{y} . Then, a simple noisy layer is used to model the relationship between true and noisy labels using these label pairs:

$$P(\tilde{y} = j | y = i) = \frac{\exp(b_{ij})}{\sum_{l=1}^k \exp(b_{il})}, \quad (2.2)$$

where

$$b_{ij} = \log \left(\frac{\sum_{t=1}^{|\mathbb{C}|} \mathbf{1}_{\{y_t=i\}} \mathbf{1}_{\{\tilde{y}_t=j\}}}{\sum_{t=1}^{|\mathbb{C}|} \mathbf{1}_{\{y_t=i\}}} \right). \quad (2.3)$$

Lange et al. (2019) further extend this method by taking the input features into consideration. That is, they first cluster contextual token vectors, and then build different distributions for each cluster, i.e., $P(\tilde{y} = j|y = i; x)$. Experimental results show that this method improves the F_1 score up to 36% over methods without noise handling when evaluate on low-resource NER settings.

2.2.2 Active learning

Active learning is a promising approach for efficient annotation, based on the hypothesis that the learning algorithm can perform better with less training if it is allowed to choose the data from which it learns (Settles, 2009). It can be used when expert annotators are available during the development cycle, but the number of instances they can annotate under budget is far less than the usual number of labelled instances needed to train a supervised model, to reach satisfactory performance. Instead of asking annotators to annotate a set of randomly sampled (*passive*) instances, *active* learning uses algorithms to choose a small set of *informative* instances to annotate.

A series of events in active learning is shown in Figure 2.2. They are repeated until the annotation budget has run out or the model performance has reached the satisfactory level. At the beginning, a model that may be trained on a small number of labelled instances or transferred from other tasks is available to make predictions on unlabelled data. The active learner chooses a small number of instances, which are considered most informative, and presents them to the expert annotators. After receiving human annotations, the model parameters can be either retrained from scratch using all available labelled data, or incrementally updated by training only on the newest batch of labelled data (Shen et al., 2018).

Although many variants exist during each step of the active learning cycle (Settles, 2009), the key component in active learning is assessing how *informative* each unlabelled instance is. In the following, we describe two widely used approaches with sequence models, including *uncertainty sampling* and *query-by-committee*, and refer the reader to (Settles and Craven, 2008; Settles, 2009; Olsson, 2009) for more options.

Uncertainty sampling. Active learner employing uncertainty-based functions chooses the instance whose label is most uncertain given the existing model.

Culotta and McCallum (2005) choose the instance for which the existing model has the least confidence in its best prediction:

$$\phi(\mathbf{x}) = 1 - P(\mathbf{y}^*|\mathbf{x}; \boldsymbol{\theta}), \quad (2.4)$$

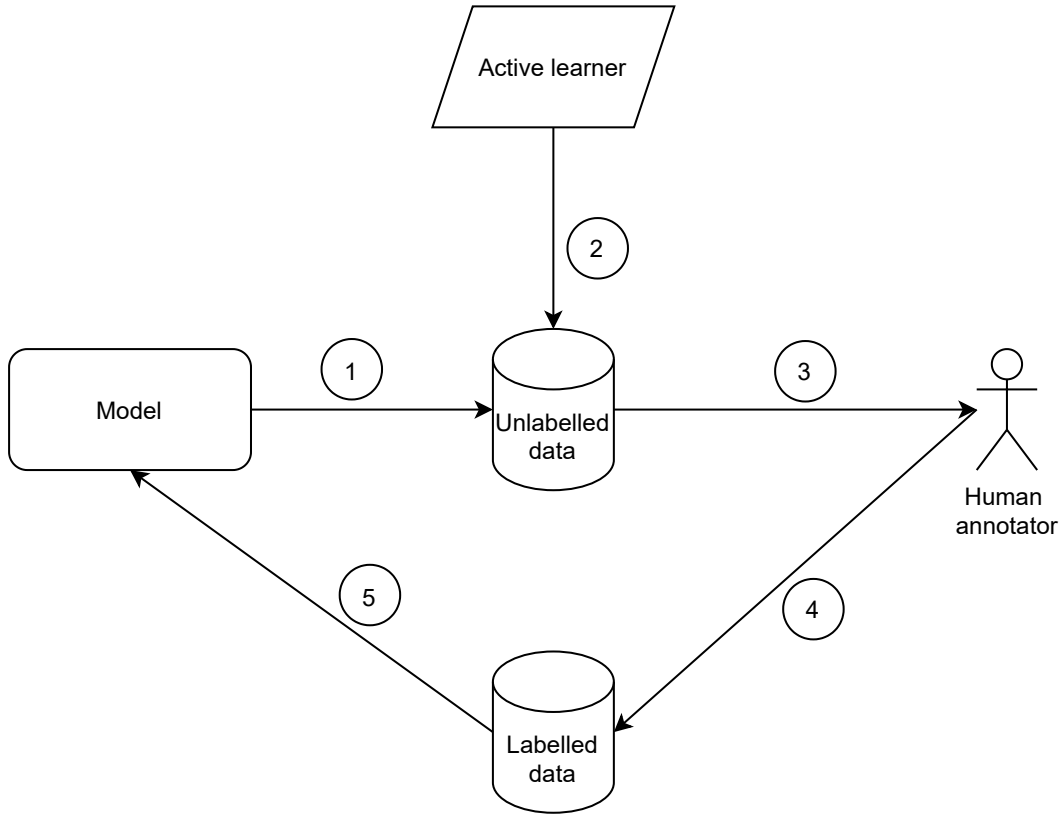


FIGURE 2.2. The active learning process usually have multiple rounds, each of which consists of five steps: (1) applying model on unlabelled data; (2) querying on unlabelled data; (3) presenting informative instances; (4) annotating instances; and (5) updating the model.

where y^* is the most likely label sequence, and θ represents the existing model.

Scheffer et al. (2001) choose the instance with the smallest margin between its two best predicted label sequences:

$$\phi(\mathbf{x}) = -(P(\mathbf{y}_1^*|\mathbf{x}; \theta) - P(\mathbf{y}_2^*|\mathbf{x}; \theta)), \quad (2.5)$$

where \mathbf{y}_1^* and \mathbf{y}_2^* are the first and second most likely labelling, respectively.

Query-By-Committee. In contrast to uncertainty sampling where only one model is used, methods belonging to query-by-committee category use multiple models, known as a committee of models. The active learner chooses the instance over which a committee of models are in most disagreement. Note that the committee needs to be comprised of diverse models.

Settles and Craven (2008) use the bagging technique to train different models. Several subsets are first randomly sampled with replacement from the original labelled training set. The same base model is then trained on each subset to create a committee of diverse models. Similarly, Shen et al. (2018) draw a committee of models via applying independently sampled dropout masks—thus different subsets of the neural network—to the same CNN-LSTM model.

To measure disagreement among a set of C models, Argamon-Engelson and Dagan (1999) introduce a measure called vote entropy:

$$\phi(\mathbf{x}) = -\frac{1}{T} \sum_{t=1}^T \sum_{j=1}^J \frac{F(\mathbf{y}_t, j)}{C} \log \frac{F(\mathbf{y}_t, j)}{C}, \quad (2.6)$$

where \mathbf{y}_t be a list of C labels predicted by all the committee models at sequence position t , and $F(\mathbf{y}_t, j)$ is the frequency of label j in \mathbf{y}_t . Shen et al. (2018) first find the most popular choice \mathbf{y}_t^* in \mathbf{y}_t , and then measure the disagreement by calculating the ratio of models which disagree with \mathbf{y}_t^* :

$$\phi(\mathbf{x}) = -\frac{1}{T} \sum_{t=1}^T \left(1 - \frac{|c : \mathbf{y}_t^{(c)} \neq \mathbf{y}_t^*|}{C}\right), \quad (2.7)$$

where $|\cdot|$ denotes cardinality of a set.

Instead of measure disagreement on the token-level, Settles and Craven (2008) describe two sequence-level measures, which consider the label sequence as a whole. Given the posterior probability of a label sequence based on a particular model, $P(\hat{\mathbf{y}}|\mathbf{x}; \boldsymbol{\theta}^{(c)})$, they first calculate the probability given the committee of models via:

$$P(\hat{\mathbf{y}}|\mathbf{x}; C) = \frac{1}{C} \sum_{c=1}^C P(\hat{\mathbf{y}}|\mathbf{x}; \boldsymbol{\theta}^{(c)}). \quad (2.8)$$

Then a set of predicted label sequences, \mathcal{N}^C , is obtained by taking the union of the N -best predictions from all models. Finally, the disagreement can be measured by calculating sequence Kullback-Leibler:

$$\phi(\mathbf{x}) = \frac{1}{C} \sum_{c=1}^C \sum_{\hat{\mathbf{y}} \in \mathcal{N}^C} P(\hat{\mathbf{y}}|\mathbf{x}; \boldsymbol{\theta}^{(c)}) \log \frac{P(\hat{\mathbf{y}}|\mathbf{x}; \boldsymbol{\theta}^{(c)})}{P(\hat{\mathbf{y}}|\mathbf{x}; C)}, \quad (2.9)$$

or sequence entropy:

$$\phi(\mathbf{x}) = - \sum_{\hat{\mathbf{y}} \in \mathcal{N}^C} P(\hat{\mathbf{y}}|\mathbf{x}; C) \log P(\hat{\mathbf{y}}|\mathbf{x}; C). \quad (2.10)$$

2.2.3 Data augmentation

Data augmentation, expanding the training set by transforming training instances without changing their labels, is heavily studied in the field of computer vision (Shorten and Khoshgoftaar, 2019). Simple augmentations, such as cropping, resizing, rotating and flipping, have become standard practices in vision tasks. However, data augmentation is still under exploration in NLP. In this section, we survey data augmentations for sentence level NLP tasks, such as text classification, natural language inference and machine translation, and group them into four categories based on how they generate augmented instances: (1) word replacement, (2) mention replacement, (3) word position swapping, and (4) using generative models.

2.2.3.1 Word replacement

Various word replacement approaches have been explored to generate augmented instances for text classification tasks. Zhang et al. (2015b) generate augmented instances by replacing words in the original instance with their synonyms, which are retrieved from an English thesaurus—WORDNET (Miller et al., 1990). They first extract all replaceable words from the original instance, and randomly choose n —determined by a geometric distribution—of them to be replaced. Then a random synonym given a word is chosen to replace the original word. Similarly, Wei and Zou (2019) randomly choose n words that are not stop words and replace each of them with one of its synonyms chosen at random. Wei and Zou show that, when the number of original training instances is small (i.e., 500), randomly choosing and replacing 10% of words from the sentence can increase the classification accuracy by 2% on average. However, when replacing too many words, for example more than 20% of words in the sentence, performance gain diminishes.

Kobayashi (2018) proposes context-aware augmentation that replace words with other words which are predicted by a language model at the word positions. Specifically, the author

pre-trains a BiLSTM language model on WIKITEXT-103 (Merity et al., 2016) – a subset of English Wikipedia articles. Then, given the surrounding words, denoted as S , at each word position i , replacement is sampled from an annealed distribution, $P(\cdot|S)^{1/\tau}$, using the language model. The parameter τ is used to control the strength of the language model. That is, when τ becomes infinity, the words are sampled from a uniform distribution. When it becomes zero, the augmentation word is always the one with the highest probability. One problem of context-aware augmentation is that the predicted word may not be compatible with the original label. For example, in a sentiment analysis dataset, the original instance ‘*the actors are fantastic*’ is labelled as *positive*. Given the word position of ‘*fantastic*’, the language model often assigns high probabilities to words such as ‘*bad*’ or ‘*terrible*’. To solve this problem, Kobayashi concatenates the embedded label y with surrounding words and use it as input to the BiLSTM language model. In other words, when training the model, Kobayashi calculate a label-conditional language model: $P(\cdot|y, S)$ instead of $P(\cdot|S)$. Evaluation results on several classification datasets show that context-aware augmentation slightly outperforms synonym-based augmentation, by accuracy of 0.5% on average.

For machine translation, word replacement has also been used to generate augmented parallel sentence pairs. Wang et al. (2018a) replace words in both the source and the target sentence by other words uniformly sampled from the source and the target vocabularies. Fadaee et al. (2017) search for contexts where a common word can be replaced by a low-frequency word, relying on recurrent language models. Similarly, Gao et al. (2019) use a monolingual language model to obtain the replacement for a randomly chosen word. Instead of predicting a single replacement word, they propose to replace the word by a soft word, which is a probabilistic distribution over the vocabulary, represented using a weighted sum of the corresponding word vectors. Experimental results show that, on both low-resource and high-resource machine translation datasets, the soft data augmentation can achieve more than 1.0 BLEU score improvement over the baseline without using data augmentation.

Document

The 1992 United States presidential election was the 52nd quadrennial presidential election, held on Tuesday, November 3, 1992. Democratic Governor Bill Clinton of Arkansas defeated incumbent Republican President George H. W. Bush, independent businessman Ross Perot of Texas, and a number of minor candidates.

Question

Who was elected the President of the United States in 1992?

Answer

Bill Clinton

FIGURE 2.3. The extractive question answering model tends to use the question type (e.g., Who) and select the spans whose nature agrees with the question type (e.g., ‘Bill Clinton’, ‘George H. W. Bush’, and ‘Ross Perot’), without the necessity to understand the question.

2.2.3.2 Mention replacement

Instead of creating augmented instances by replacing individual words, replacement can be employed at the mention level, usually with the help of an external knowledge base and heuristic rules.

After observing that question answering models tend to astray by selecting a text span that shares the answer’s type but has the wrong underlying entity (Figure 2.3), Raiman and Miller (2017) design an augmentation strategy to make the model more robust to surface form variation. It includes three steps:

- (1) Extract nominal groups in the training set using a part of speech tagger.
- (2) Perform string matching with entities in Wikidata.
- (3) Randomly replace matched entities in the training set with other entities of the same category in Wikidata.

Specifically, they extract 47,598 entities in SQUAD that fall under 6,380 Wikidata *instance of* types. During each training epoch, T —a hyperparameter, tuned from a range $[0, 10^5]$ —augmented instances are generated and used in combination with the original training set.

Experimental results on SQUAD show that the proposed data augmentation improves the performance by F_1 of 1.0.

In order to remove gender bias from coreference resolution systems, Zhao et al. (2018) propose to generate an augmented set where all male entities are replaced by female entities, and vice versa, and train the model on both original and augmented sets. They use a rule based approach consisting of two steps. First, named entities are anonymised. For example, ‘*John went to his house*’ would be anonymised to ‘*EI went to his house*’. Then a dictionary of gendered terms and their realisation as the opposite gender is used to change all matching tokens. For example, ‘*she*’ is changed to ‘*he*’, ‘*Mr.*’ is changed to ‘*Mrs.*’. Finally, the augmented instance ‘*EI went to her house*’ is generated and added to the training set. Evaluation results on a benchmark dataset focused on gender bias show that this data augmentation can effectively remove the gender bias without significantly affecting the model performance on other coreference benchmark datasets.

2.2.3.3 Word position swapping

Wei and Zou (2019) randomly choose two words in the sentence and swap their positions to augment text classification training sets. They use only one parameter to control the number of words changed based on the sentence length. Experimental results show that random swap can yield high performance gains when less than 20% of words in the sentence are swapped, but decline when more than 30% of words are swapped.

Min et al. (2020) explore syntactic transformations (e.g., subject/object inversion, passivisation) to augment the training data for Natural Language Inference (NLI) to mitigate over-fitting. This transformation does not attempt to ensure the naturalness of the generated examples, neither the correctness of labels. For example, in the subject/object inversion transformation, the sentence ‘*The carriage made a lot of noise*’ is transformed into ‘*A lot of noise made the carriage*’, and the gold label of the augmented instance is set to *neutral* if the original label is *entailment*. Experimental results show that the proposed augmentation does not harm overall performance on the MNLI test set, but it can help the model achieve better generalisation, evaluated on HANS.

2.2.3.4 Generative models

Instead of creating an augmented instance by manipulating one or several tokens in the original instance, some approaches aim to create a new instance via generative models.

Yu et al. (2018) train a question answering model with data augmented by back-translation from a neural machine translation model. Specifically, they use two translation models, one model from English to French and another model from French to English. They feed the document from an original instance into the English-to-French model to obtain k French translations via the decoder using beam search. Then each of the French translation is passed through a French-to-English model with beam decoder, and can thus obtain k^2 paraphrased instances in total. Experimental results on SQUAD show that the proposed data augmentation can improve the performance by F_1 of 1.1, when the training data is made three times as large by adding augmented instances.

Similarly, Xia et al. (2019b) convert data from a high-resource language to a low-resource language, using a bilingual dictionary and an unsupervised machine translation model in order to expand the machine translation training set for the low-resource language. Results show that, under extreme low resource settings, the proposed data augmentation can improve translation quality measured by BLEU compared to supervised back-translation baselines.

2.2.4 Summary

In this section, we reviewed three promising approaches—distant supervision, active learning and data augmentation—that aim to achieve high accuracy with as little annotating efforts as possible. Different approaches make use of different types of resources (Table 2.4), and therefore can be suitable for different scenarios. For example, active learning requires expert-in-the-loop, and distant supervision makes use of domain-specific knowledge base or domain knowledge for designing heuristic rules. They are good options once these resources are available. In contrast, data augmentation is the most flexible approach, since some augmentation methods can be applied without the requirement of any domain-specific resources, e.g., word replacement. Encouraged by its adaptability and existing data augmentation methods

	Labelled data	Unlabelled data	Knowledge base	Domain expert
Active learning		✓		✓
Data augmentation	✓			
Distant supervision		✓	✓	✓

TABLE 2.4. Requirement of different types of resources by each approach.

for sentence-level NLP tasks, we investigate easy to use data augmentation methods for NER, which will be detailed in Chapter 3.

We note that there are other approaches to overcome the low resource problem, such as unsupervised learning (Collins and Singer, 1999; Etzioni et al., 2005; Zhang and Elhadad, 2013), as well as transfer learning, which we describe in the following section. These approaches are not mutually exclusive, therefore we can combine them. For example, transfer learning—pre-training language representation models on unlabelled data, and then fine-tuning on target labelled data—has become a standard practice in NLP. Methods belonging to other approaches can be combined with transfer learning, such as using off-the-shelf pre-trained models as the baseline model.

2.3 Transfer Learning

The standard supervised learning requires sufficient labelled data to train a decent performing model, given a particular task, domain and language. In other words, each model is trained individually for a combination of task, domain and language. In contrast, transfer learning explores the relatedness between tasks, domains and languages. The knowledge gained in solving a *source* task in a *source* domain and a *source* language is applied to solve the *target* task in the *target* domain and *target* language (Ruder, 2019).

Yang et al. (2017) develop a transfer learning approach for sequence tagging and design different neural architectures for cross-domain, cross-task, and cross-lingual transfer settings. In the cross-domain transfer, the authors share all parameters of the model—BiLSTM-CRF—and perform a label mapping on top of the classifier (Figure 2.4a). Note that, cross-domain transfer typically has mappable label sets that labels in different domains can be mapped to

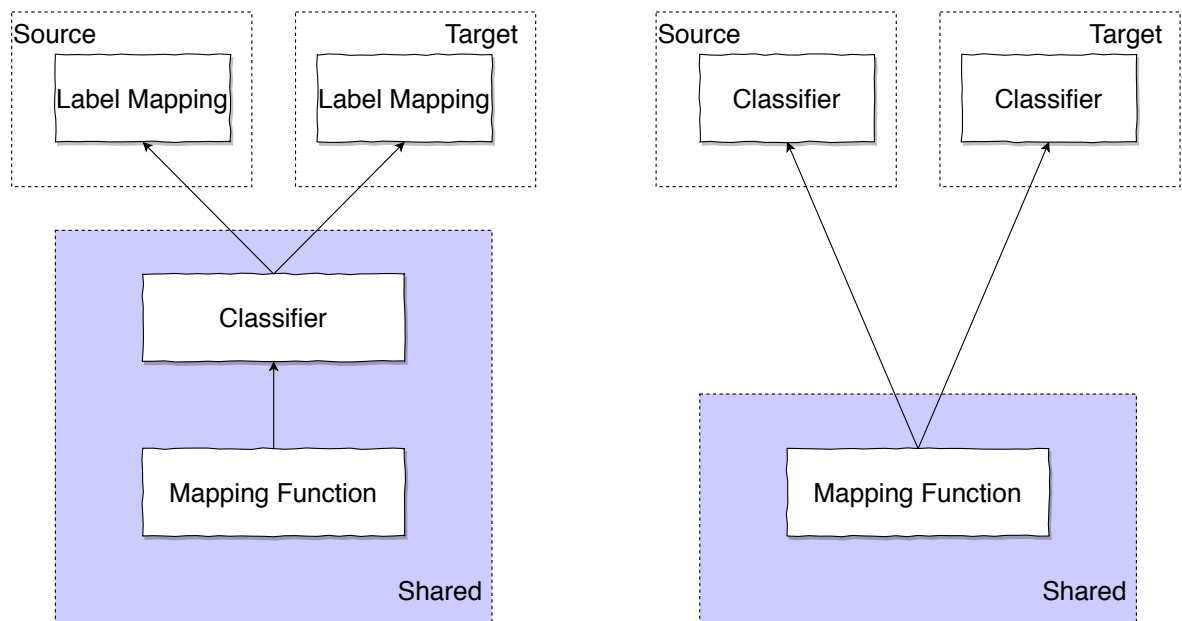
each other. For the unmappable setting, Yang et al. consider it the same as cross-task transfer, and each task learns a separate classifier (Figure 2.4b). The cross-lingual transfer is achieved by exploiting the morphologies shared by different languages. For example, the morphological similarity between ‘*Canada*’ in English and ‘*Canadá*’ in Spanish can be exploited for NER. The transfer learning architecture shares only the character level mapping function, which takes a sequence of characters as input, building a token feature vector (Figure 2.4c).

Although improvements have been reported by using cross-domain, cross-task, and cross-lingual transfers, a big challenge in these approaches is finding related source task, domain, and language. In other words, the knowledge learned in solving a source task in a source domain and language can be transferred, only if the knowledge is indeed shared between the source and target. In this thesis, we focus on English datasets; therefore, we do not discuss cross-lingual in details; we refer readers to (Rahimi et al., 2019; Ruder et al., 2019b; Conneau et al., 2020) for more discussions.

2.3.1 Cross-task transfer

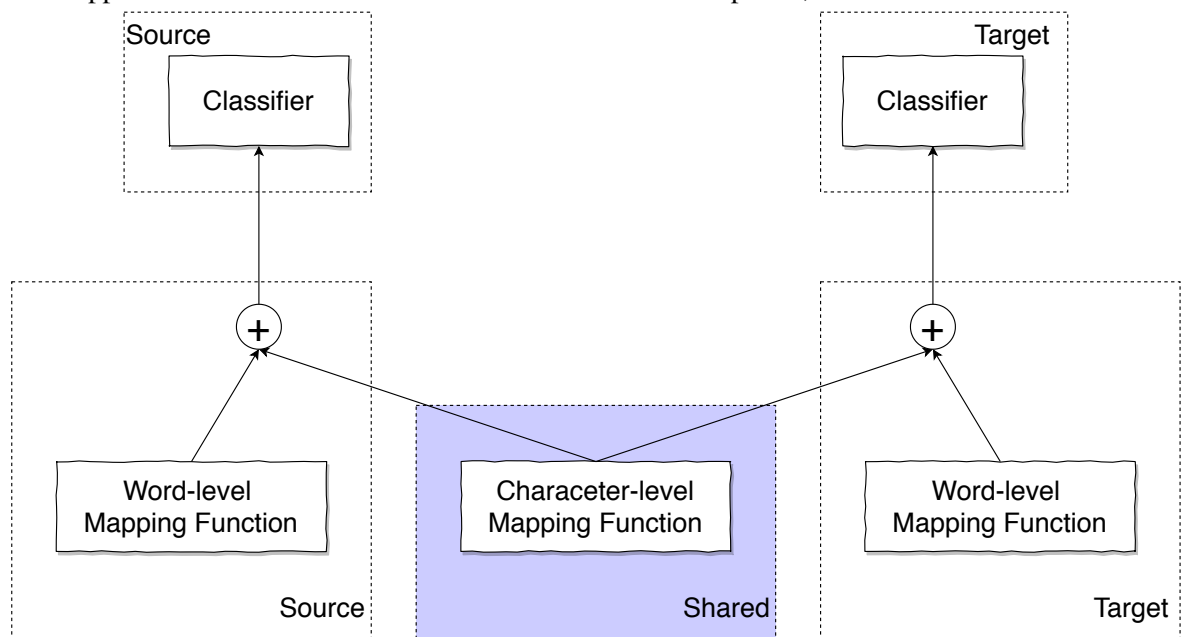
Patra and Moniz (2019) point out that it is easier and faster for an annotator to answer a yes/no question than to recognise all entity mentions. That is, the cognitive load of selecting whether an entity mention is present or not in the sentence is less than that of highlighting and annotating mentions with their entity categories. Therefore, they propose to use a model which is trained on a sentence level multi-label classification task—whether an entity mention is present or not—and transfer it to the entity recognition task. Evaluation results on CONLL 2003 show that the proposed method works surprisingly well, achieving F_1 score of 81.1.

Ruder et al. (2019a) propose a meta-architecture for multi-task learning. They use part-of-speech tagging—a fundamental syntactic task—as the auxiliary task, and observe that chunking, NER, and semantic role labelling tasks can benefit from the auxiliary task, outperforming the single task learning baseline. Similar ideas have also been explored by Collobert et al. (2011); Søgaard and Goldberg (2016). For example, Søgaard and Goldberg (2016) use *low level* NLP task, such as part-of-speech tagging, to improve the *higher level* tasks, such as



(A) Cross-domain transfer when the label sets are mappable.

(B) Cross-domain transfer when the label sets are disparate, and cross-task transfer.



(C) Cross-lingual transfer. The \oplus also represents a neural network that takes both character-level and word-level feature vectors as input and creates the final token feature vector.

FIGURE 2.4. Neural architectures for the settings of cross-domain, cross-task, cross-lingual transfer proposed in (Yang et al., 2017).

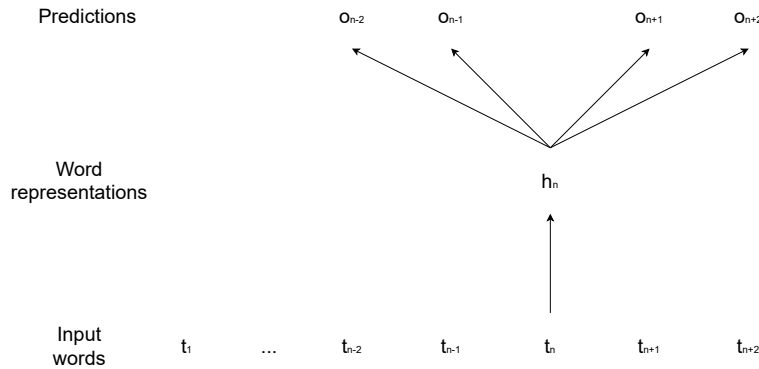


FIGURE 2.5. The Skip-gram model aims to learn word representations that can be used to predict the surrounding words.

chunking and CCG Supertagging. They design specialised multi level LSTM networks that have part-of-speech supervision at the innermost layer, and other tasks the outermost layer.

In contrast to transferring from a source task that requires labelled data, transfer learning techniques can make the most of limited labelled data by incorporating language representation models pre-trained on a large amount of unlabelled data (Mikolov et al., 2013a; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019). Many pre-training tasks have demonstrated their effectiveness for different downstream tasks. In this section, we briefly review three pre-training tasks and refer readers to (Wang et al., 2019) for more options.

Skip-gram model. The Skip-gram model, introduced by (Mikolov et al., 2013a), is an efficient method for learning static word representations from unlabelled text. The training objective of the Skip-gram model is to build word representations that can be used to predict the surrounding words in a sentence (Figure 2.5). Given a sequence of tokens $\{t_i\}_{i=1}^N$, the Skip-gram model aims to maximise the objective:

$$\frac{1}{N} \sum_{i=1}^N \sum_{-c \leq j \leq c, j \neq 0} \log p(t_{i+j} | t_i),$$

where c , a hyper-parameter, is the size of the context window.

Skip-gram model has been shown to effectively capture syntactic and semantic information of words (Mikolov et al., 2013b). However, its main disadvantage is that it always assigns the

same vector to the word, no matter what the context of the word is. The word representations learned using Skip-gram model are therefore called *static* word representations.

Masked language modelling. In contrast to *static* word representations, where one word is always assigned the same vector, *contextual* word representations can assign different vectors to the same word, depending on its context.

Dai and Le (2015) explored the idea of pre-training recurrent language model and transferring it to the downstream supervised models. They use unlabelled data from Amazon reviews to pre-train the language model and find that it can improve classification accuracy on the Rotten Tomatoes dataset. Peters et al. (2017) extend the single direction language model to bidirectional. Based on these efforts, Devlin et al. (2019) propose the masked language modelling pre-training task to better capture contexts from both sides. Different to Peters et al. (2018) who build two language models—left-to-right and right-to-left—which are trained separately, the masked language modelling is a fill-in-the-blank task. That is, a small set of tokens are masked, and the model needs to use the context tokens to try to predict what the masked tokens should be (Figure 2.6).

Replace token detection. The *replace token detection* task, proposed by Clark et al. (2020), is a sample efficient variant of masked language modelling (Figure 2.7). Instead of replacing some tokens as the special [MASK] token, Clark et al. (2020) employ a small generator network to generate plausible alternatives. Then the discriminator network predicts whether each token in the input is replaced by a generator sample or not.

Another difference between the masked language modelling and replace token detection pre-training tasks is that the former is performed on only masked tokens, whereas the latter is defined over all input tokens. Therefore, a replace token detection pre-training task requires less compute, measured using floating point operations. Clark et al. (2020) show that it performs comparably to masked language modelling pre-training task while using less than 1/4 of their compute and outperforms masked language modelling when using the same amount of compute. Because pre-training a language representation model using replace

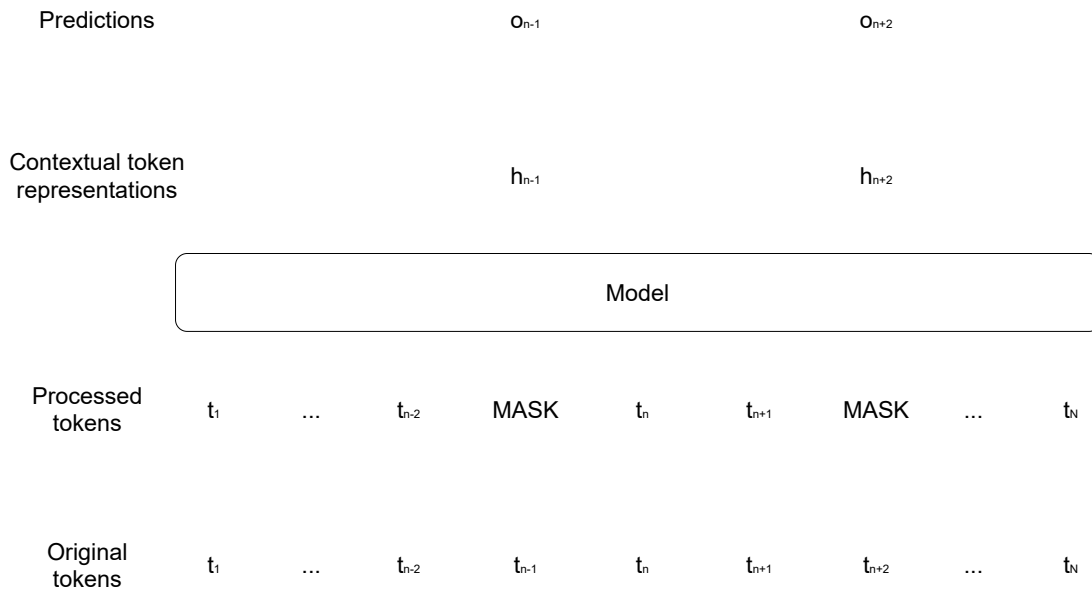


FIGURE 2.6. The mask language modelling pre-training task aims to learn contextual word representations that can be used to predict what the masked token is.

token detection task can be done within several days using a single GPU, we thus use it for our investigation in Chapter 4.

2.3.2 Cross-domain transfer

Han and Eisenstein (2019) propose *domain-adaptive fine-tuning*, in which the language representation models are adapted by masked language modelling on text from the target domain. They evaluate this approach on sequence labelling in two challenge domains: Early Modern English and Twitter. Results show that domain-adaptive fine-tuning yields substantial improvements over strong BERT baselines, with particularly strong results on out-of-vocabulary words. Similarly, Gururangan et al. (2020) investigate whether it is helpful to tailor a pre-trained model to the domain of a target task. They show that domain-adaptive pre-training—continue pre-training on a large corpus of unlabelled domain-specific text—leads to performance gains. Moreover, task-adaptive pre-training—continue pre-training on the unlabelled set for a given task—improves performance even after domain-adaptive pre-training. Gururangan et al. (2020) consider domain vocabularies containing the top 10K most

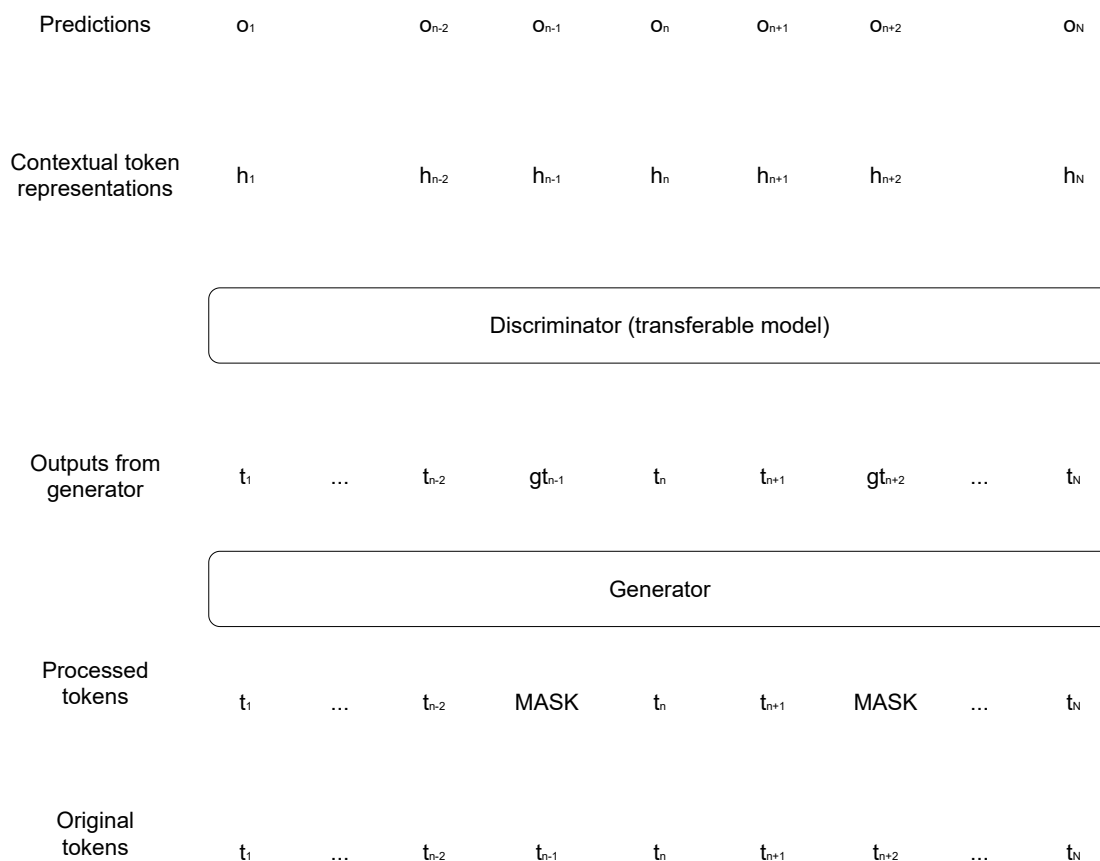


FIGURE 2.7. The replace token detection task aims to train the discriminator to predict whether the token is the original token or a fictional token.

frequent uni-grams and use the vocabulary overlap as the measure of domain similarity. They find that pre-training data used in RoBERTa—over 160GB of uncompressed text, consisting of Wikipedia, books, stories, news articles, and web content extracted from URLs shared on Reddit—have a very low vocabulary overlap with datasets sampled from biomedical scholarly articles (27.3%) and computer science scholarly articles (19.2%).

Moore and Lewis (2010) propose a cross-entropy difference selection method to select in-domain training data to build auxiliary language models for use in tasks such as machine translation and speech recognition. Given the target data set \mathcal{T} and a generic source \mathcal{S} , they aim to select a subset of the available source data as language model training data. Let $H_T(s)$ be the per-word cross-entropy, according to a language model trained on \mathcal{T} , of a sentence s drawn from \mathcal{S} . Let $H_S(s)$ be the per-word cross-entropy of s according to a language

model trained on a random sample of \mathcal{S} . Moore and Lewis (2010) score the sentences from \mathcal{S} according to $H_T(s) - H_S(s)$, and all sentences whose score is less than a threshold are selected as in-domain training data. A similar idea was explored by Klakow (2000), who estimates a language model from the entire \mathcal{S} , and scores the subset of \mathcal{S} by the change in the log likelihood of \mathcal{T} according to another language model, where that subset is removed from training data. Those subsets whose removal would decrease the log likelihood of \mathcal{T} more than a threshold are selected.

Plank and van Noord (2011) evaluate measures of domain similarity and their impact on dependency parsing accuracy. Given a target article to parse and a collection of annotated articles, they want to select the most similar articles to train the parser which is then evaluated on the target article. Both probabilistically motivated similarity functions—such as Jensen-Shannon divergence, and skew divergence—and geometrically motivated distance functions—such as cosine, euclidean, and variational distance functions—are evaluated on different features. Plank and van Noord (2011) find that comparing article topic distributions estimated by Latent Dirichlet Allocation (LDA) (Blei et al., 2003) using variational distance function or Jensen-Shannon divergence can effectively find the most similar source, and using these automatic measures can outperform using human annotated genre labels. In addition to above mentioned similarity measures, Asch and Daelemans (2010) explore to use Rényi entropy and Bhattacharyya coefficient to estimate the impact of domain similarity on cross-domain transfer.

2.3.3 Summary

Inspired by these efforts that use domain similarity to nominate suitable data for labelling or training statistical language models. We explore whether these similarity measures can also be used to nominate in-domain data for pre-training large scale neural language representation models.

Our work is also inspired by several lines of work that aim to link the known to the unknown, studying its impact. Ramponi and Plank (2020) study the implications of variations of

language on model performance. They argue that treating text as just input data to machine learning is problematic, and it is important to study how covert and overt factors, such as genre, social-demographic aspect, stylistic and data sampling strategy, impact results, and take these factors into consideration in modelling and evaluation. Johnson et al. (2018) predict a system’s accuracy using larger training data from its performance on much smaller pilot data. In Chapter 4, we aim to link the similarity between source pre-training data and target task data to the effectiveness of pre-trained models. In other words, we aim to design a cost-effective approach that predicts the usefulness of pre-trained models for target datasets based on the similarity between the source pre-training data and the target task data.

2.4 Complex Entity Recognition

Commonly, the NER problem is framed as: given a sequence of tokens, output a list of spans, each of which is an entity mention in text. Recall that a span is a consecutive sequence of tokens, or an individual token. The mention can therefore be represented using the starting and ending indices of the span: I_s , I_e . Additionally, each mention is assigned to an entity category. This perspective imposes two constraints:

- (1) An entity mention consists of a continuous sequence of tokens, where all the tokens indexed between I_s and I_e are part of the entity mention; and,
- (2) These linear spans do not overlap with each other. In other words, no token can belong to more than one entity mention.

Most of the existing NER datasets in the generic domain, for example CONLL 2003 (Sang and Meulder, 2003), or ONTONOTES 5.0 (Weischedel et al., 2011), are annotated satisfying these two constraints. Therefore, conventional sequence taggers achieve state-of-the-art effectiveness in these datasets (Lample et al., 2016; Ma and Hovy, 2016; Akbik et al., 2018; Baevski et al., 2019).

However, in practice, there are domains, such as the biomedical domain, in which there can be entity mentions nested, overlapping, and discontinuous (see examples in Figure 2.8).



- a) ... activation of the HIV-1 enhancer following ...
- 
- DNA: HIV-1 enhancer
Virus: HIV-1
- b) ... had intense pelvic and back pain ...
- 
- ADE: intense pelvic pain
ADE: back pain

FIGURE 2.8. Examples involving nested, overlapping and discontinuous entity mentions. In (a), ‘*HIV-1 enhancer*’ and ‘*HIV-1*’ are nested entity mentions. In (b), ‘*intense pelvic pain*’ and ‘*back pain*’ overlap, and ‘*intense pelvic pain*’ is a discontinuous mention.

These *complex entity mentions* cannot be directly recognised by conventional sequence taggers because they break the above mentioned constraints based on which sequence tagging techniques are built.

In this section, we first describe these complex entity mentions in details (Section 2.4.1). We then review the existing methods which are proposed to recognise complex entity mentions and categorise them into token-level (Section 2.4.2), span-level (Section 2.4.3), and sentence-level (Section 2.4.4) approaches. Finally, we identify the research gap, that our proposed method (described in Chapter 5) is going to fill.

2.4.1 Definitions of complex entity mentions

Nested entity mentions. One entity mention is completely contained by the other. We call both of the mentions involved nested entity mentions. Figure 2.8 a) is an example taken from the GENIA corpus (Kim et al., 2003). Here, ‘*HIV-1 enhancer*’ is a DNA mention, and it contains another mention ‘*HIV-1*’, which is a virus.

Multi-type entity mentions. An extreme case of nested entity mentions is one in which a span corresponds to multiple mentions. For example, in the EPPI corpus (Alex et al., 2007), proteins can also be annotated as drug/compound, indicating that the protein is used as a

drug to affect the function of a cell. Such a mention should be classified as both protein and drug/compound. In this case, we consider this mention as two mentions of different categories, and these two mentions contain each other.

Overlapping entity mentions. Two entity mentions overlap, but neither is completely contained by the other. Figure 2.8 b) is an example taken from the CADEC corpus (Karimi et al., 2015a), which is annotated for adverse drug events (ADE) and relevant concepts. In this example, two ADEs: ‘*intense pelvic pain*’ and ‘*back pain*’, share a common token ‘*pain*’, and neither is contained by the other.

Discontinuous entity mentions. The mention consists of a discontinuous sequence of tokens. In other words, the mention contains at least one interval. In Figure 2.8 b), ‘*intense pelvic pain*’ is a discontinuous entity mention since it is interrupted by ‘*and back*’.

Recognising complex entity mentions is important because these mentions can hold very useful information (Ringland et al., 2019). First, the nested and overlapping structures themselves are already good indicators of the relationship between different entities involved. For example, an ORGANISATION mention ‘*University of Sydney*’ contains a LOCATION mention ‘*Sydney*’. This structure implies the location of the organisation, and recognition of these mentions can potentially speed up the construction of a knowledge base (Ringland et al., 2019). Second, recognising complex entity mentions can simplify the design of downstream tasks. For example, separating overlapping mentions rather than identifying them as a single mention is important for a downstream entity linking task, where the assumption is that the input mention refers to one entity, and the task can thus be regarded as one-to-one mapping (Shen et al., 2014). Third, recognising complex entity mentions can improve the performance of other NLP tasks. For example, entity mentions often have fixed representations in different languages. Therefore, recognising entity mentions, especially those discontinuous entity mentions, can improve the performance of a machine translation system (Klementiev and Roth, 2006). Last but not least, we notice that similar complex structures also exist in other NLP tasks, such as multi-word expressions recognition (Baldwin and Kim, 2010; Rohanian et al., 2019) and constituency parsing (Coavoux et al., 2019; Coavoux and Cohen,

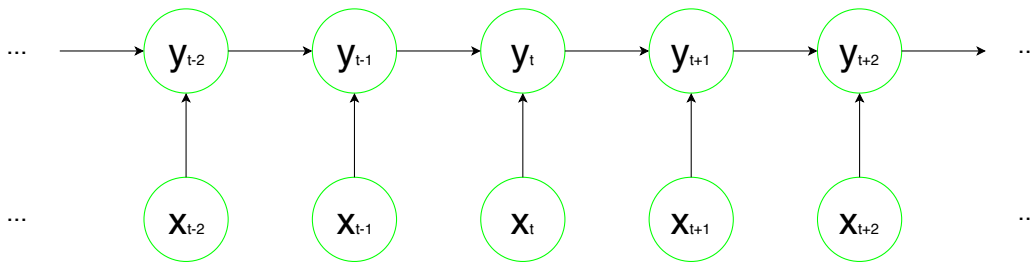


FIGURE 2.9. In a linear-chain CRF model, the output for each token depends on the representation of that token in context and the output for the previous token.

2019). We believe the ideas proposed for recognising complex entity mentions should also apply to similar complex structures in other tasks.

2.4.2 Token-level Approach

The main component of sequence tagging techniques is a structural prediction model, which takes a sequence of contextual token representations as input and outputs a tag for each token. Figure 2.9 is an illustration of such a model. That is, in a linear-chain CRF model, the tag of one token depends on both the token representation and the tag of the previous token. These local decisions are then chained together to perform joint inference, and the tag sequence predicted by the tagger is finally decoded into entity mentions using explicit rules. We categorise the methods based on conventional sequence tagging as token-level approach.

In vanilla sequence tagging models, the intermediate outputs for each token are usually BIO schema tags. However, since the BIO tags cannot effectively represent complex entity mentions, the first natural direction is to expand the BIO tag set so that different kinds of complex entity mentions can be captured. Metke-Jimenez and Karimi (2016) introduce a BIO variant schema to represent discontinuous and overlapping entity mentions. That is, in addition to the BIO prefixes, four new position indicators, BD, ID, BH, and IH are proposed to denote **B**eginning of **D**iscontinuous body, **I**nside of **D**iscontinuous body, **B**eginning of **H**ead, and **I**nside of **H**ead. Here, the token sequences which are shared by multiple mentions are called head, and the remaining parts of the mention are called body. Figure 2.10 is an encoding example using this schema. ‘*pain*’ is the beginning of the head that is shared by

had intense pelvic and back pain .
 O BD ID O B BH O

FIGURE 2.10. An encoding example of two adverse drug event mentions: ‘*intense pelvic pain*’ and ‘*back pain*’.

two mentions, and therefore tagged as *BH*. ‘*intense pelvic*’ is the body of a discontinuous mention, while ‘*back*’ is the beginning of a continuous mention. Here, we keep only the position indicator and remove the entity category ‘ADE’, since this schema can only represent overlapping mentions of the same entity category. Note that, even in this simple example, it is still impossible to represent several mentions unambiguously. For example, this encoding can also be decoded as having three mentions: ‘*intense pelvic pain*’, ‘*back pain*’ and ‘*pain*’.

Tag variants are also proposed to deal with complex structures with specialised constraints. To deal with nested NER (one mention is completely contained by the other mention), Alex et al. (2007) propose a *joined labelling* variant that each token is assigned a tag by concatenating the tags of all levels of nesting. For example, the token ‘*HIV-1*’ in Figure 2.8 is assigned a tag ‘B-DNA+B-Virus’, indicating that the token is the beginning token within a DNA mention and the beginning token within a VIRUS mention. Then the tagger is trained on the data containing the joined labels. During the inference stage, the joined labels are decoded into their original BIO format for each entity category. Rohanian et al. (2019) introduce BIOG tags for discontinuous structure without overlapping involved. The new G position indicator is used for tokens in between the components. Muis and Lu (2017) propose to assign tags to the gaps between tokens, while still regarding the problem as a sequence labelling problem. In other words, they model the mention boundaries instead of the role of tokens in forming mentions (Figure 2.11).

Instead of elaborating schema to encode entity mentions with complex structures, another direction based on sequence tagging techniques is to employ multiple sequence tagging models or layers, that are arranged in a series. Alex et al. (2007) employ several sequence tagging models, each of which is used to recognise entity mentions belonging to a group of several entity categories without nested structures. Similarly, Ju et al. (2018) stack several BiLSTM-CRF layers together, each of which is used to recognise entity mentions belonging

... directed by the [IL2]-regulatory-region] or by ...

X X X S EC C E X X

FIGURE 2.11. An example of mention separators encoding two nested entity mentions: ‘*IL2*’ and ‘*IL2 regulatory region*’. Muis and Lu (2017) design three mention separators: S, also denoted as [, indicating a mention is starting at the next token; E (]), indicating a mention is ending at the previous token; and C (-), indicating a mention is continuing to the next token. X means none of the three separators applies. The standard sequence tagger, which takes as input a sequence of N tokens and outputs a sequence of $N-1$ mention separators, can be used to recognise nested NER.

to a particular nesting layer. Note that, although these two methods achieve decent results in nested NER benchmarks, they have some difficulties in dealing with special nested structures. The cascade approach proposed by Alex et al. (2007) cannot deal with nested mentions of the same entity category. For example, one DNA mention might contain another DNA mention. The layered approach proposed by Ju et al. (2018) cannot deal with multi-type entity mentions. For example, one mention might be annotated as both PROTEIN and DRUG/COMPOUND.

2.4.3 Span-based Approach

The vanilla idea of span-based approach enumerates all possible spans – up to a certain length in a sentence – as potential entity mentions. It then determines whether a span is a valid entity and what is its entity category. These candidate spans do not need to exclude each other, so the predicted entity mentions can also overlap with each other. This advantage makes a span-based approach a strong option for nested NER, and it has been extensively investigated.

Given vector representations of each token h_i in the sentence $\mathbf{H} = h_1, \dots, h_i, \dots, h_n$ and a candidate span (i, j) , the key decision of span-based approaches is how to build the span representation and score the span for each entity category. Once all candidate spans are represented as fix-length vectors and scored for each entity category, they are ranked by the scores and the top-ranked spans are outputted as the final predictions.

A summary of representative techniques to build span representations and to score span for each entity category is shown in Table 2.5. Building span representations via directly

using boundary token representations as well as tokens within the span is the simplest solution. Sohrab and Miwa (2018) represent the span by concatenating the boundary token representations and the average of all token representations within the span ($\frac{1}{j-i+1} \sum_{k=i}^j \mathbf{h}_k$). The span representation is then passed to a softmax output layer to classify the span into a specific entity category or non-entity. Similarly, Luan et al. (2018, 2019) construct span representation by concatenating the boundary token representations (\mathbf{h}_i and \mathbf{h}_j), an attention-based soft *headword*, and embedded span width features, and then use a feed-forward neural network to produce per-class scores for span.

Xu et al. (2017) employ Fixed-sized Ordinarily Forgetting Encoding (FOFE) to encode the span and its contexts into a fixed-size vector and then use a feed-forward neural network to predict its entity category. FOFE mimics bag-of-words but incorporates a forgetting factor (α in Table 2.5) to capture positional information (Zhang et al., 2015a). Xu et al. create both word-level and character-level features for each span and its left and right contexts: FOFE code of the span ($f(i, j)$); FOFE code for left context including the span ($f(1, j)$), FOFE code for left context excluding the span ($f(1, i - 1)$); FOFE code for right context including the span ($f(i, n)$); FOFE code for right context excluding the span ($f(j + 1, n)$).

Yu et al. (2020) argue the contexts of the start and end of the span are different. They apply two separate feed-forward neural networks to create different boundary representations (\mathbf{h}_s and \mathbf{h}_e , in Table 2.5) for the start and end of the span. Then they use a biaffine model (Dozat and Manning, 2016) to score the span. Xia et al. (2019a) run an additional BiLSTM on top of the token representations and use an attention mechanism to let tokens within the span attend to contexts to get the span representation. Finally, a two-layer feed-forward neural network is used to score the span.

One shortcoming of span-based approaches is that all candidate spans are scored independently. The exhaustive enumeration of possible spans creates a large number of negative instances. That is, the majority of candidate spans belong to a non-entity category. Also, interactions among mentions are not explored, because all span representations are built in parallel on top of the same underlying token representations.

Model	Representing and scoring spans
(Xu et al., 2017)	$\mathbf{h}(i, j) = \begin{bmatrix} f(1, i-1) \\ f(1, j) \\ f(i, j) \\ f(i, n) \\ f(j+1, n) \end{bmatrix}$ <p>where $f(i, j) = \begin{cases} \mathbf{h}_i & \text{if } i = j \\ \alpha \cdot f(i, j-1) + \mathbf{h}_j & \text{otherwise} \end{cases}$</p> $\text{score}(i, j) = \text{SOFTMAX}(\mathbf{W} \cdot \mathbf{h}(i, j))$
(Luan et al., 2018, 2019)	$\mathbf{h}(i, j) = \begin{bmatrix} \mathbf{h}_i \\ \mathbf{h}_j \\ \text{SELF ATTENTION}(\mathbf{H}) \\ \text{SPAN WIDTH FEATURE} \end{bmatrix}$ $\text{score}(i, j) = \mathbf{W} \cdot \mathbf{h}(i, j)$
(Sohrab and Miwa, 2018)	$\mathbf{h}(i, j) = \begin{bmatrix} \mathbf{h}_i \\ \frac{1}{j-i+1} \sum_{k=i}^j \mathbf{h}_k \\ \mathbf{h}_j \end{bmatrix}$ $\text{score}(i, j) = \text{SOFTMAX}(\mathbf{W} \cdot \mathbf{h}(i, j))$
(Xia et al., 2019a)	$\mathbf{e} = \text{BILSTM}(\mathbf{h}_i \cdots \mathbf{h}_j)$ $\mathbf{a} = \text{SOFTMAX}(\mathbf{H}\mathbf{W}\mathbf{e}^T)$ $\mathbf{C} = \mathbf{a} \star \mathbf{H}$ $\mathbf{m} = \text{BILSTM}(\mathbf{C})$ $\mathbf{h}(i, j) = \begin{bmatrix} \mathbf{m} \\ \mathbf{e} \end{bmatrix}$ $\text{score}(i, j) = \text{SOFTMAX}\left(\mathbf{W}_2 \cdot \left(\sigma\left(\mathbf{W}_1 \cdot \mathbf{h}(i, j) + \mathbf{b}_1\right)\right) + \mathbf{b}_2\right)$
(Yu et al., 2020)	$\mathbf{h}_s = \text{FFNN}_s(\mathbf{h}_i)$ $\mathbf{h}_e = \text{FFNN}_e(\mathbf{h}_j)$ $\text{score}(i, j) = \mathbf{h}_s^\top \mathbf{W}_1 \mathbf{h}_e + \mathbf{W}_2 \cdot (\mathbf{h}_s \oplus \mathbf{h}_e) + \mathbf{b}$

TABLE 2.5. A summary of techniques to represent and score span, given a sequence of token representations $\mathbf{h}_1, \dots, \mathbf{h}_n$. $\mathbf{h}(i, j)$, being a fixed-length vector representation of the span, with its dimension being a hyper-parameter. $\text{score}(i, j)$ is the (normalised) score for the span from i to j inclusive, where $1 \leq i \leq j \leq n$. $\text{score}(i, j)$ is usually a c -dimension vector, where c is the number of entity categories, including a special category for non-entity.

We describe efforts on overcoming this problem from different perspectives:

Solving class imbalance problem. Given a sequence of n tokens, if we enumerate all possible spans in the sentence, there are in total $\frac{n \times (n+1)}{2}$ candidate spans. These candidate

spans belong to one of three categories: (1) exact match with a gold entity mention; (2) partial overlap with a gold mention; and, (3) disjoint with any mention. The latter two (negative instances) significantly outnumber the first exact match ones (positive instances). This class imbalance problem may result in low predictive accuracy.

To solve this problem, Xu et al. (2017) and Xia et al. (2019a) use a down-sampling strategy. That is, they fix the total number of candidate spans in each training batch, including all positive spans and sampled negative spans.

Sun et al. (2019) remove those negative spans that highly overlap with spans corresponding to gold mentions. The negative span b is used for training, only if

$$\max \left(\left[\text{IoU}(b, g) \text{ for } g \text{ in } G \right] \right) \leq \Gamma, \quad (2.11)$$

where G is the set of gold entity mentions. $\text{IoU}(b, g)$ is a function measuring how many tokens are shared between two spans:

$$\text{IoU}(b, g) = \frac{\text{length}(b \cap g)}{\text{length}(b \cup g)} \quad (2.12)$$

and Γ is a hyperparameter tuned on different datasets.

Reducing search space. Instead of exhaustive classifying over all possible spans, a two-stage paradigm is investigated to reduce the size of candidate mentions. Zheng et al. (2019) propose a boundary-aware model, where first sequence labelling models are used to detect possible span boundaries, and then the span based models are used to predict entity categories of a small number of candidate spans. Similarly, Xia et al. (2019a) separate the task into two stages: deciding whether the candidate span is an entity mention or not via a detector, and then classifying detected candidates into predefined entity categories via a classifier.

Lin et al. (2019) detect entity mentions by using what they call *head-driven phrase structures*. They first identify possible head words of entity mentions, and then recognise the mention boundaries by exploiting phrase structures. They argue that although entity mentions might nest with each other, they cannot share the same head words, and the head words are informative to decide the entity category. They also propose an objective function—bag loss—which

does not require gold head word annotations. This is done by exploiting the association between words and entity categories.

Modelling surrounding mentions. To take the surrounding mentions of a given span into consideration, Xu et al. (2017) introduce a *2nd-pass* mechanism. They train two models: one standard model, and the other model using outputs from the first model, where the predicted entity categories are used to replace the entity mentions. During inference, the span score is the linear interpolation between scores from these two models.

Luan et al. (2018) propose a multi-task learning framework where entity recognition, relation extraction, and coreference resolution are treated as classification problems with shared span representations. By sharing low-level LSTM encoder, information about relation types with surrounding mentions and coreferences can be used to create input span representations to entity classifier. Instead of sharing only LSTM encoder, Luan et al. (2019) further extend the multi-task model using dynamically constructed span (node) graphs. At each training step, the most confident entity spans are treated as nodes in a graph structure, and arcs are confidence-weighted relation types and coreferences. Then, the span representations are refined using updates which are propagated from neighbouring relation types and co-referred entities.

2.4.4 Sentence-level Approach

Instead of predicting whether a token belongs to an entity mention and its role in the mention (token-level approach) or whether a consecutive sequence of tokens form an entity mention (span-level approach), some methods predict directly a combination of entity mentions within a sentence. We call these methods sentence-level approach.

McDonald et al. (2005) consider NER as a structured multi-label classification. Instead of starting and ending indices, they represent each entity mention using the set of token positions that belong to the mention. An example of this representation, with each token tagged using an I/O schema is shown in Figure 2.12. The advantage of this method is that the

Bill and Hilary Clinton traveled to Canada .

P	O	O	P	O	O	O	O
O	O	P	P	O	O	O	O
O	O	O	O	O	O	L	O

FIGURE 2.12. An example of a sentence with three entity mentions: ‘*Bill Clinton*’ and ‘*Hilary Clinton*’ are PERSON mentions, and ‘*Canada*’ is a LOCATION mention. P and L refer to the entity categories: PERSON and LOCATION, respectively.

representation is very flexible as it allows entity mentions consisting of discontinuous tokens and does not require mentions to exclude each other. Using this representation, the NER problem is converted into the multi-label classification problem of finding up to k correct labels among all possible $(T + 1)^n$ labels, where k is a hyper-parameter of the model, T is the number of entity categories, and n is the length of the sentence. Note that labels do not come from a pre-defined category but depend on the sentence being processed. McDonald et al. use large-margin online learning algorithms to train the model, so that the scores of the correct labels (entity mentions) are higher than those of all other possible incorrect mentions. Another advantage of this method is that the outputs of the model are unambiguous for all kinds of complex entity mentions and easy to be decoded. However, the method suffers from a $O(n^3T)$ inference complexity.

Finkel and Manning (2009) use a discriminative constituency parser to recognise nested entity mentions. They represent each sentence as a constituency tree, where each mention corresponds to a phrase in the tree. In addition, each node needs to be annotated with its parent and grandparent labels, so that the parser can learn how entity mentions nest. Ringland (2016) also employ a joint model using the Berkeley parser (Petrov et al., 2006), and show that it performs well even without specialised NER features. However, one disadvantage of these parsing based models, as in (McDonald et al., 2005), is that their time complexity is cubic in the number of tokens in the sentence. Furthermore, the high quality parse training data, which is not always available, plays a crucial role in the success of the joint model (Li et al., 2017).

Lu and Roth (2015) propose a novel hypergraph to represent exponentially many possible nested mentions in one sentence, and one sub-hypergraph of the complete hypergraph can therefore be used to represent a combination of mentions in the sentence. The mention hypergraph consist of five types of nodes:

A^k **nodes:** represent all mentions whose left boundaries are exactly at or after the k -th token;

E^k **nodes:** represent all mentions whose left boundaries are exactly at the k -th token;

T_j^k **nodes:** represent all mentions whose left boundaries are exactly at the k -th token and have the mention type j ;

I_j^k **nodes:** represent all mentions which contain the k -th token and have the mention type j ;
and

X **nodes:** indicate the completion of a path.

Hyper-edges, each of which consists of a parent node and an ordered list of child nodes, are used to connect nodes. Figure 2.13 is an example of such a sub-hypergraph, which represents two nested entity mentions.

The training objective of this hypergraph-based model is to maximise the log-likelihood of training instances consisting of the sentence and mention-encoded hypergraph. During inference, the model first predicts a sub-hypergraph among all possible sub-hypergraphs of the complete hypergraph, and predicted mentions can be decoded from the output sub-hypergraph. Different to Lu and Roth (2015) who build hand-crafted features defined over the input sentence and the output hypergraph structure, Katiyar and Cardie (2018) learn the hypergraph representation using features extracted from a recurrent network.

Although this hypergraph-based model enjoys a time complexity that is linear in the number of tokens in the input sentence, it suffers from some degree of ambiguity during decoding stage. For example, when one mention is contained by another mention with the same entity category and their boundaries are all different, the hypergraph can be decoded in different ways. This ambiguity comes from the fact that, if one node has multiple parent nodes and multiple child nodes, there is no mechanism to decide which of the parent node is paired with which child node. Therefore, Wang and Lu (2018) propose an extension of the I node where

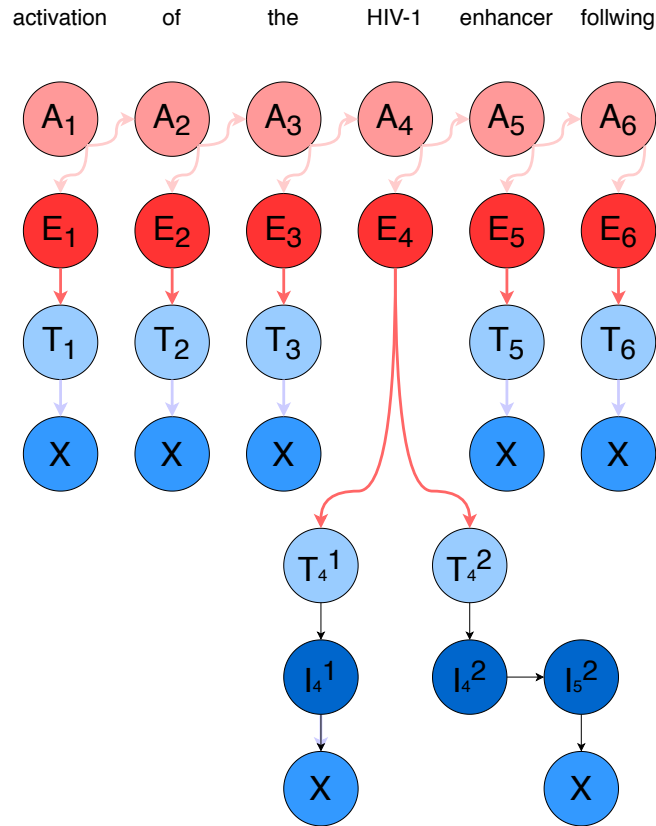


FIGURE 2.13. An example sub-hypergraph with two nested entity mentions: ‘*HIV-1*’ (VIRUS) and ‘*HIV-1 enhancer*’ (DNA). Here, one mention corresponds to a path consisting of (AETI+X) nodes. Specifically, the path (A₄E₄T₄¹I₄¹X) corresponds to the mention ‘*HIV-1*’, and the path (A₄E₄T₄²I₄²I₅²X) corresponds to the mention ‘*HIV-1 enhancer*’.

they use $I_{i,n}^k$ nodes to represent all mentions of type k which contain the j -th token and start with the i -th token.

To represent discontinuous mentions, Muis and Lu (2016) expand the node types in the hypergraph representation to capture discontinuous mentions. That is, they add two new node types: B for tokens within the mention, and O for tokens belonging to part of the gap. Figure 2.14 is an example of the sub-hypergraph, which encodes two mentions: ‘*muscle pain*’ and ‘*muscle fatigue*’. Wang and Lu (2019) propose a two-stage approach that all spans are first identified using the hypergraph representation and then a classifier is used to predict whether two spans form a discontinuous mention.

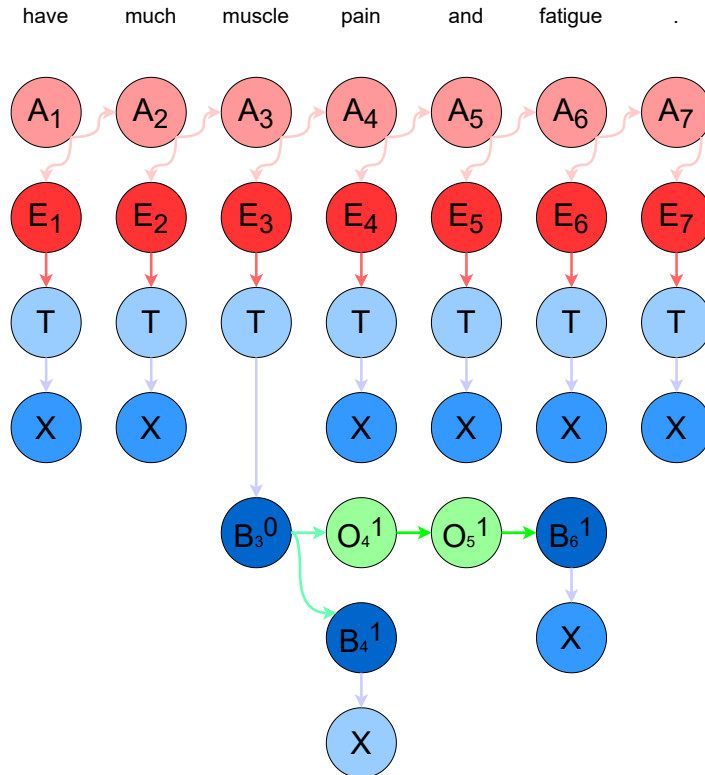


FIGURE 2.14. An example sub-hypergraph with two entity mentions: ‘*muscle pain*’ and ‘*muscle fatigue*’. Muis and Lu (2016) extend the hypergraph representation proposed by Lu and Roth (2015) to capture discontinuous mentions through two new node types: B_k^i representing the k -th token is part of the i -th component of an entity mention, and O_k^i representing the k -th token appears in between $(i - 1)$ -th and i -th components of an entity mention. In this example, the path $(A_3E_3TB_3^0B_4^1X)$ corresponds to the mention ‘*muscle pain*’ and the path $(A_3E_3TB_3^0O_4^1O_5^1B_6^1X)$ corresponds to the discontinuous mention ‘*muscle fatigue*’.

2.4.5 Summary

Despite the potential applications of complex NER recognition, there is comparatively few studies on recognising discontinuous and overlapping mentions. The span-based approach focuses on building effective span representations which are used to predict whether the span is an entity mention and its entity category. However, it cannot be directed used for discontinuous NER, because discontinuous mentions consist of multiple spans.

We observe that there is usually a trade-of between the expressiveness power and the modelling difficulty. In other words, the more flexible (less constraints) the representation is, the more

interactions are ignored, therefore the model might be more difficult to train. For example, the multi-label representation proposed by McDonald et al. (2005) is the most flexible representation (Figure 2.12); however, it does not take any interactions between different mentions into consideration. Tang et al. (2018) empirically show that a less flexible representation—BIO variant tagging based model—can outperform the multi-label representation when the training data are limited.

Methods belonging to token-level and sentence-level approaches first predict intermediate representations, and then the intermediate representations are decoded into entity mentions. That is, token level representations use a sequence of tags (Tang et al., 2013a; Metke-Jimenez and Karimi, 2016) as the intermediate representation, and sentence-level approach uses a graph structure (Lu and Roth, 2015; Katiyar and Cardie, 2018). Inspired by these approaches, in Chapter 5, we propose a transition-based model for discontinuous NER. Similar to token-level and sentence-level approaches, our transition-based model first predicts the intermediate representation, i.e., a sequence of actions, but greatly reduce the ambiguity problem in these two approaches. Our method also employs effective methods to build span representations, which are inspired by span-based approach.

Data Augmentation for NER

Data augmentation, expanding the training set by transforming training instances without changing their labels, is heavily studied in the field of computer vision and sentence level NLP tasks. Inspired by these efforts, we design several easy to use data augmentation methods for NER. Through experiments on two English datasets from the biomedical domain, we demonstrate that our proposed augmentation methods can boost performance over a strong baseline where large scale pre-trained models are used, especially when the original labelled training set is small.

3.1 Overview

Modern deep learning techniques typically require a large amount of labelled data for training (Bowman et al., 2015; Conneau et al., 2017). However, in real world applications, such large labelled data sets are not always available. This is especially true in some specific domains, such as the biomedical domain, where annotating data requires expert knowledge and is usually time-consuming (Karimi et al., 2015a; Nye et al., 2018).

Different approaches have been investigated to solve this *low-resource* problem (Hedderich et al., 2020). For example, transfer learning pre-trains language representations on self-supervised or rich-resource *source* tasks and then adapts these representations to the *target* task (Ruder, 2019; Gururangan et al., 2020). Data augmentation expands the training set by applying transformations to training instances without changing their original labels (Shorten and Khoshgoftaar, 2019).

Recently, there has been an increased interest on applying data augmentation techniques on sentence-level NLP tasks, such as text classification (Wei and Zou, 2019; Xie et al., 2019), natural language inference (Min et al., 2020), and machine translation (Wang et al., 2018a). Augmentation methods explored for these tasks include creating augmented instances by manipulating a few words in the original instance, such as word replacement (Zhang et al., 2015b; Wang and Yang, 2015; Cai et al., 2020), random deletion (Wei and Zou, 2019), and word position swap (Şahin and Steedman, 2019; Min et al., 2020); or creating entirely artificial instances via generative models, such as variational autoencoders (Yoo et al., 2019; Mesbah et al., 2019) and back-translation models (Yu et al., 2018; Iyyer et al., 2018).

Different from these sentence-level NLP tasks, NER is usually regarded as a token-level NLP task. That is, for each token in the sentence, an NER model predicts a label indicating whether the token belongs to an entity mention and which entity category the mention belongs to. Therefore, applying transformations to individual tokens may also change their labels. Due to such a difficulty, data augmentation for NER is relatively less studied. In this chapter, we describe our efforts to fill this gap by exploring data augmentation techniques for NER, solved as a sequence tagging problem.

3.2 Proposed Data Augmentation Methods

We surveyed the existing data augmentation techniques for sentence-level NLP tasks in Section 2.2. Inspired by these efforts, we design several easy to use data augmentation methods for NER. Note that our proposed methods do not rely on any external trained models, such as machine translation models (Yu et al., 2018; Iyyer et al., 2018) or syntactic parsing models (Şahin and Steedman, 2019), which are by themselves difficult to train in low-resource domain specific scenarios.

Given an original training instance, consisting of a sequence of tokens and the corresponding sequence of labels, we use the following transformations to create augmented instances.

Method	Instance							
None	She	did	not	complain	of	headache	or	
	O	O	O	O	O	B-problem	O	
LwTR	<i>L.</i>	<i>One</i>	not	complain	of	headache	<i>he</i>	
	O	O	O	O	O	B-problem	O	
SR	any	other	neurological	symptoms	.			
	B-problem	I-problem	I-problem	I-problem	O			
MR	<i>She</i>	did	<i>non</i>	complain	of	headache	or	
	O	O	O	O	O	B-problem	O	
SiS	<i>whatsoever</i>	<i>former</i>	neurologic	symptom	.			
	B-problem	I-problem	I-problem	I-problem	O			
MR	She	did	not	complain	of	<i>neuropathic pain</i>		
	O	O	O	O	O	<i>B-problem I-problem</i>		
SiS	<i>syndrome</i>	or	<i>acute</i>	<i>pulmonary disease</i>	.			
	<i>I-problem</i>	O	<i>B-problem</i>	<i>I-problem I-problem</i>	O			
SiS	<i>not</i>	<i>complain</i>	<i>She</i>	<i>did</i>	<i>of</i>	headache	or	
	O	O	O	O	O	B-problem	O	
SiS	<i>neurological</i>	<i>any</i>	<i>symptoms</i>	<i>other</i>	.			
	B-problem	I-problem	I-problem	I-problem	O			

TABLE 3.1. An original training instance and different types of augmented instances. We highlight changes using *italics*.

Label-wise Token Replacement (LwTR). For each token which is not a stop word, we use a binomial distribution to randomly decide whether it should be replaced. If yes, we then use a label-wise token distribution, built from the original training set, to randomly select another token with the same label. Thus, we keep the original label sequence unchanged. Taking the instance in Table 3.1 as an example, there are five tokens replaced by other tokens which share the same label as the original tokens.

Synonym Replacement (SR). Our second approach is similar to LwTR, except that we replace the token with one of its synonyms retrieved from WORDNET (Miller et al., 1990).

Note that the retrieved synonym may consist of more than one token. Its BIO labels can be derived using a straightforward rule: If the replaced token is the first token within a mention (i.e., the corresponding label is ‘B-Entity’), we assign the same label to the first token of the retrieved multi-word synonym, and ‘I-Entity’ to the other tokens.

Mention Replacement (MR). For each mention in the instance, we use a binomial distribution to randomly decide whether it should be replaced. If yes, we randomly select another mention from the original training set which has the same entity category as the replacement. The corresponding sequence of BIO labels can be changed accordingly. For example, in Table 3.1, the mention ‘*headache* [B-problem]’ is replaced by another problem mention ‘*neuropathic pain syndrome* [B-problem I-problem I-problem]’.

Shuffle within Segments (SiS). We first split the token sequence into segments of the same entity category. Thus, each segment corresponds to either an entity mention or a sequence of tokens that does not belong to any mention. For example, the original instance in Table 3.1 is split into five segments: ‘*She did not complain of* [Out-of-Mention]’, ‘*headache* [Problem]’, ‘*or* [Out-of-Mention]’, ‘*any other neurological symptoms* [Problem]’, ‘. [Out-of-Mention]’. Then for each segment, we use a binomial distribution to randomly decide whether it should be shuffled. If yes, the order of the tokens within the segment is shuffled, while the label order is kept unchanged.

All. We also explore the augmentation of the training set using all aforementioned augmentation methods. That is, for each training instance, we create multiple augmented instances, one per transformation.

	I2B2-2010			NCBI-DISEASE		
	Train	Dev	Test	Train	Dev	Test
# Sentences	13,868	2,447	27,625	5,424	923	940
# Tokens	129,087	20,454	267,249	135,701	23,969	24,497
# Mentions	14,376	2,143	31,161	5,134	787	960

TABLE 3.2. The descriptive statistics of the two English datasets from the biomedical domain: I2B2-2010 (Uzuner et al., 2011) and NCBI-DISEASE (Doğan et al., 2014).

3.3 Evaluation

We present an empirical analysis of the data augmentation methods described in Section 3.2 on two English datasets from the biomedical domain¹: I2B2-2010 (Uzuner et al., 2011) and NCBI-DISEASE (Doğan et al., 2014).

We use a BERT-CRF model (Beltagy et al., 2019; Baevski et al., 2019) as the backbone model, and we investigate the impact of applying data augmentation on training data of different sizes.

3.3.1 Datasets

I2B2-2010 focuses on the identification of three entity types of problem, treatment and test from patient reports. We use the train-test split from its corresponding shared task and randomly select 15% of sentences from the training set as the development set. NCBI-DISEASE contains scholarly articles annotated with disease names. We use the train-dev-test split provided by the authors. Descriptive statistics of these two datasets is listed in Table 3.2.

To simulate a low-resource setting, we select the first 50, 150, 500 sentences which contain at least one mention from the complete training set to create the corresponding small, medium, and large subsets (denoted as S, M, L in Table 3.3, whereas the complete training set is denoted as F). Note that we apply data augmentation only on the training set, without changing the development and test sets.

¹In (Dai and Adel, 2020), we also evaluate these methods on a dataset from the materials science domain.

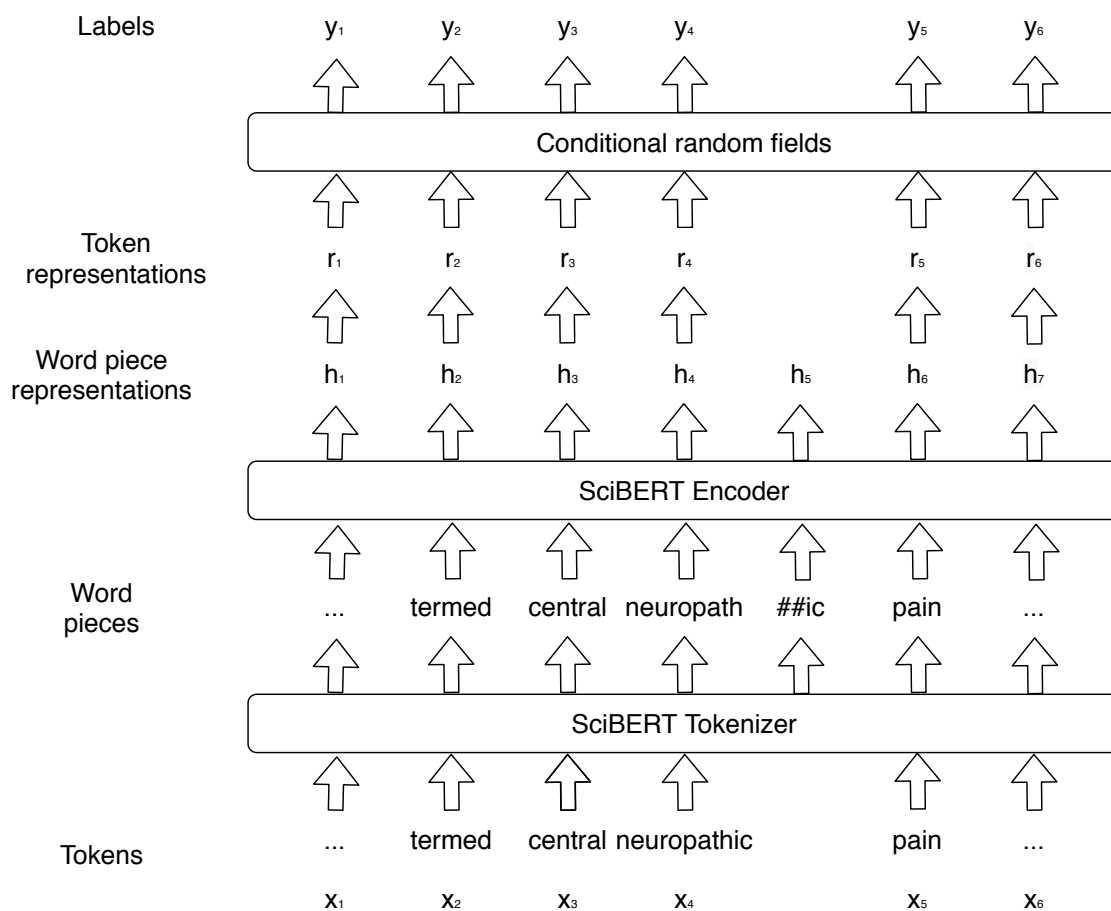


FIGURE 3.1. High level overview of the BERT-CRF model.

3.3.2 Backbone model

We regard the NER task as a token-level sequence tagging problem, where each token in the sentence is assigned a tag. The tag can be used to infer whether the token is the first token within a mention, inside a mention or does not belong to any mention.

The backbone model, illustrated in Figure 3.1, is a BERT-CRF model (Beltagy et al., 2019; Baeovski et al., 2019). It takes advantage of large scale pre-trained language models—using BERT-based encoder to create contextual representations for each token, and a probabilistic graphical model—using conditional random fields (Sutton and McCallum, 2007) to capture dependencies between neighbouring tags.

BERT-based encoder. Given a sentence, the tokenizer, coupled with the pre-trained BERT-based model, first converts each token in the sentence into word pieces. That is, if the original token does not exist in the vocabulary, it will be segmented into several pieces from the vocabulary (Sennrich et al., 2015). Then the word pieces are mapped to dense vectors—token embeddings—via a lookup table. Finally, the sum of token embeddings and positional embeddings, which indicate the position of each token in the sequence, are fed into a stack of multi-head self-attention and fully-connected feed-forward layers (Vaswani et al., 2017). Following the study in (Devlin et al., 2019), we use the final outputs corresponding to the first word piece within each token as the token representation.

Recent studies on domain-specific BERT models show that effectiveness on downstream tasks can be improved when the BERT models are further pre-trained on in-domain data (Gururangan et al., 2020). We thus choose SciBERT (Beltagy et al., 2019), which is pre-trained on full text of scholarly articles about biology and computer science, and fine-tune it on the target NER task. In our preliminary experiments, we observe that SciBERT achieves significant better results than vanilla BERT (Devlin et al., 2019) and slightly better results than BioBERT (Lee et al., 2020).

Conditional random fields (CRF). Instead of assigning a tag to each token independently, we model them jointly using a conditional random fields. That is, given a sequence of token representations $\mathbf{R} = (r_1, r_2, \dots, r_n)$, we aim to predict a sequence of tags $\mathbf{y} = (y_1, y_2, \dots, y_n)$ which has the maximum probability over all possible tag sequences. This conditional probability can be calculated using:

$$p(\mathbf{y} \mid \mathbf{R}) = \frac{e^{s(\mathbf{R}, \mathbf{y})}}{\sum_{\tilde{\mathbf{y}} \in \mathbf{Y}_{\mathbf{R}}} e^{s(\mathbf{R}, \tilde{\mathbf{y}})}}$$

and

$$s(\mathbf{R}, \mathbf{y}) = \sum_{i=0}^n A_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i},$$

where $A_{i,j}$ is the compatibility score of a transition from the tag i to tag j , and $P_{i,j}$ is the score of the tag j given token representation r_i .

Corpus	Size	Baseline	LwTR	SR	MR	Sis	All
i2b2-2010	S	34.6 ± 1.4	39.9 ± 0.7	43.8 ± 2.0	39.6 ± 1.3	39.2 ± 1.8	42.6 ± 1.4
	M	62.9 ± 1.0	64.0 ± 1.3	64.5 ± 0.5	63.5 ± 0.8	63.1 ± 1.4	63.9 ± 1.8
	L	69.6 ± 0.3	70.5 ± 1.6	70.7 ± 1.1	70.6 ± 0.9	71.0 ± 1.2	70.5 ± 1.0
	F	87.6 ± 0.3	87.3 ± 0.2	87.7 ± 0.2	87.6 ± 0.1	87.1 ± 0.2	86.8 ± 0.3
NCBI-disease	S	59.9 ± 2.6	62.9 ± 1.4	63.6 ± 2.3	65.1 ± 1.3	63.0 ± 1.1	63.8 ± 1.3
	M	71.6 ± 1.4	73.2 ± 1.5	74.7 ± 1.0	73.6 ± 1.1	73.4 ± 1.1	73.3 ± 1.2
	L	81.0 ± 0.4	80.4 ± 1.1	82.2 ± 1.0	80.5 ± 1.3	80.6 ± 1.0	81.3 ± 0.5
	F	87.6 ± 0.3	85.7 ± 1.1	87.9 ± 0.7	88.1 ± 0.7	87.4 ± 0.4	86.0 ± 1.0
Δ			1.1	2.5	1.7	1.2	1.6

TABLE 3.3. Evaluation results in terms of span-level F_1 score. Small set contains 50 training instances; Medium contains 150 instances; Large contains 500 instances; Full uses the complete training set. Results that are better than the baseline model without using data augmentation are highlighted in bold. underline: the result is significantly better than the baseline model without data augmentation (paired student’s t-test, p: 0.05)

The parameters, of both the SciBERT encoder and CRFs, are trained jointly to maximise the conditional probability of gold tag sequence given the training sentences.

3.3.3 Experimental results

The evaluation results on the effectiveness of data augmentation methods are shown in Table 3.3. We use the Micro-average string match F_1 score to evaluate the effectiveness of the models. The model which is most effective on the development set, measured using the F_1 score, is finally evaluated on the test set. All experiments are repeated five times with different random seeds. Mean values and standard deviations are reported. The Δ row shows the averaged improvement due to data augmentation, comparing against the baseline without using data augmentation. In general, we find that all data augmentation methods improve over the baseline, and synonym replacement outperforms other augmentation on average.

Another observation is that the data augmentation methods are more effective when the original training sets are small. For example, all data augmentation methods achieve improvements when the training set contains only 50 training instances. In contrast, when the complete training sets are used, only synonym replacement and mention replacement achieve

improvements. This has also been observed in previous work on applying data augmentation on other NLP tasks (Fadaee et al., 2017; Şahin and Steedman, 2019; Xia et al., 2019b).

3.4 Analysis

After demonstrating the effectiveness of proposed data augmentation methods, we present an analysis of the best two performing transformations: synonym replacement and mention replacement. We aim to provide practical suggestions on hyperparameter settings as well as understandings about how they improve the performance.

3.4.1 The impact of hyperparameters

For each augmentation method, we tune the number of augmented instances per original training instance from a list of numbers: {1, 3, 6, 10}. We also tune the p value of the binomial distribution which is used to decide whether a token or a mention should be replaced. It is searched over the range from 0.1 to 0.7, with an incremental step of 0.2. We perform grid search to find the best combination of these two hyperparameters on the development set.

The main question we aim to answer is how much augmentation is enough? More specifically, how the number of augmented instances per original training instance affects performance, and how the ratio a token, or a mention, is replaced affects performance?

Figure 3.2 shows the impact of the number of augmented instances per original training instance on the performance gain of synonym replacement and mention replacement. We use the improvement of absolute F_1 score over the baseline without using data augmentation as the performance gain.

In general, we find that *larger number of augmented instances can bring larger performance gain, especially when the training sets are small (i.e., 50 training instances)*. However, the performance gain becomes relatively small when the number of augmented instances per original training instance is greater than 6. The second observation is that when the training sets are medium (i.e., 150 training instances) or large (i.e., 500 training instances), the benefits

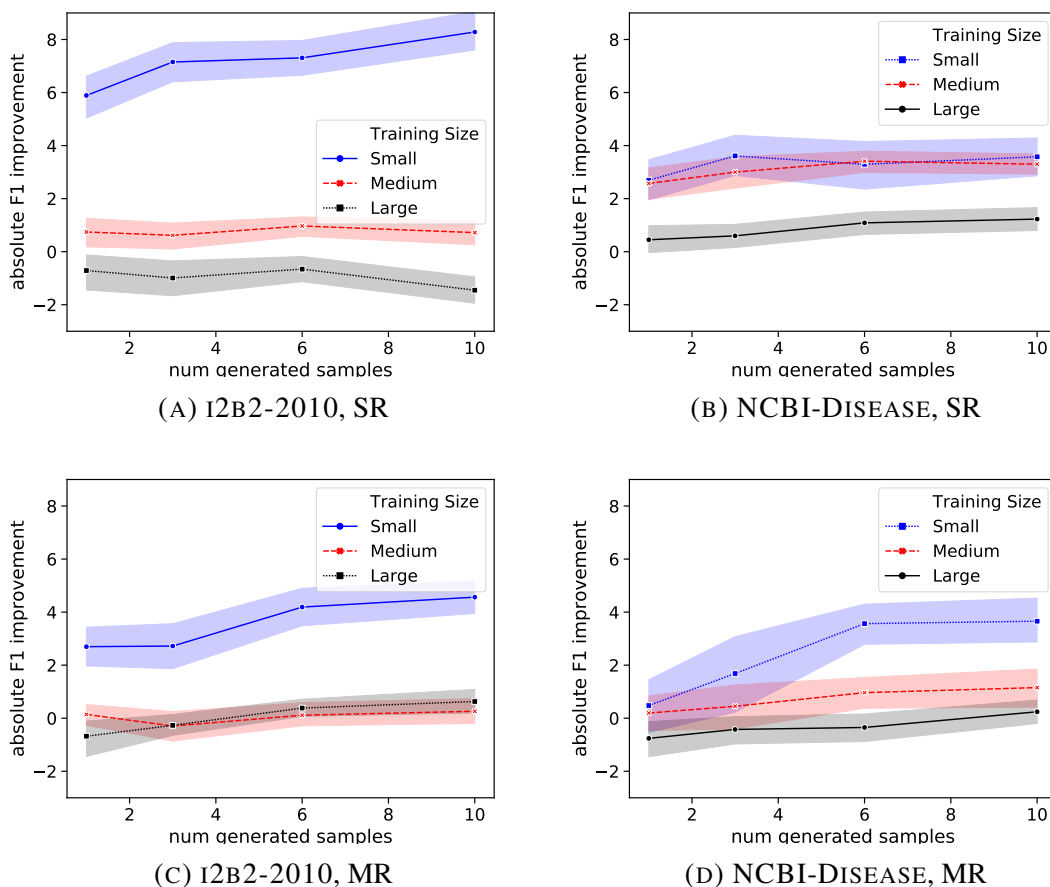


FIGURE 3.2. Impact of the number of augmented instances per original training instance on the effectiveness of data augmentation. SR: synonym replacement. MR: mention replacement.

of more augmented instances become small. On i2B2-2010, creating more augmented instances using synonym replacement on large training set even decreases the performance.

Figure 3.3 A and B shows the impact of the ratio a token is replaced with one of its synonyms on the performance gain. We note a moderate ratio (e.g., 0.3 or 0.6) performs well across different setups. If the ratio is too small, the augmented instances may be very similar to the original one. Training on these augmented instances may have a similar effect as training on the original training instances for more epochs. In contrast, a large ratio is more likely to create syntactically invalid instance. These syntactically invalid instances are noisy, and training on such a combination of small amount of clean data and large amount of noisy data may underperform training on clean data only (more discussions in Section 2.2.1).

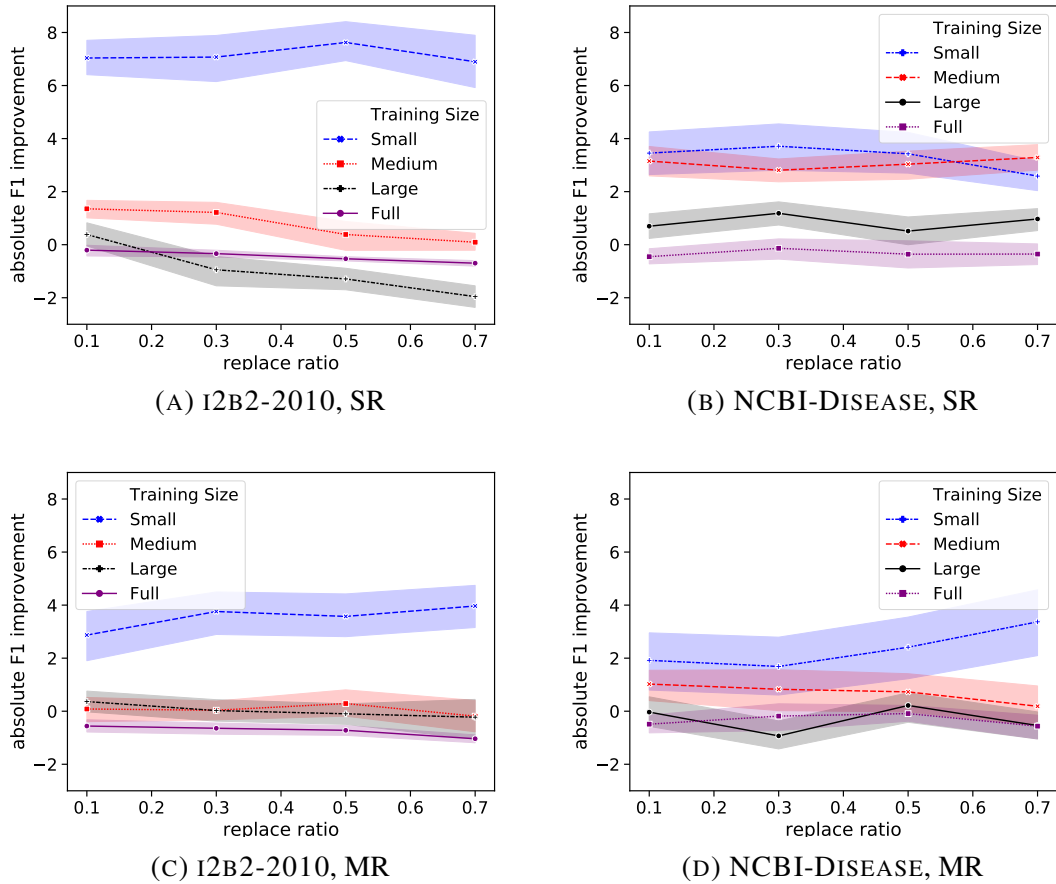


FIGURE 3.3. The impact of the ratio a token or a mention is replaced on the effectiveness of data augmentation. SR: synonym replacement. MR: mention replacement.

The pattern with mention replacement is different from the one with synonyms replacement (Figure 3.3 C and D). When the training sets are small, increasing the ratio a mention is replaced with another mention of the same entity category always enlarges the performance gain. We find that most of these entity mentions have similar part of speech patterns. That is, most of them are either nouns or noun phrases. Because of this feature, the risk of creating syntactically invalid instance is much lower using mention replacement than using synonyms replacement. It can be the reason why *a moderate ratio for synonyms replacement is best, whereas mention replacement can benefit from a large replace ratio.*

	i2B2-2010		NCBI-DISEASE	
	FPs	FNs	FPs	FNs
Baseline \cap SR \cap MR	12574 (26.9)	13500 (68.8)	170 (20.5)	197 (45.8)
Baseline \cap SR \cap \neg MR	2424 (5.2)	2001 (10.2)	53 (6.4)	50 (11.6)
Baseline \cap \neg SR \cap MR	4261 (9.1)	658 (3.4)	105 (12.7)	18 (4.2)
Baseline \cap \neg SR \cap \neg MR	6674 (14.3)	1613 (8.2)	122 (14.7)	58 (13.5)
\neg Baseline \cap SR \cap MR	2949 (6.3)	335 (1.7)	72 (8.7)	42 (9.8)
\neg Baseline \cap SR \cap \neg MR	5034 (10.8)	1107 (5.6)	84 (10.1)	46 (10.7)
\neg Baseline \cap \neg SR \cap MR	12767 (27.3)	421 (2.1)	223 (26.9)	19 (4.4)

TABLE 3.4. The comparison of different types of errors—FPs (false positives) and FNs (false negatives)—made by the baseline model without using data augmentation and models using Synonym Replacement (SR) and Mention Replacement (MR) data augmentation methods. \cap indicates the intersection of two sets, and \neg indicates the negative set. For example, the ‘FPs’ column corresponding to the ‘Baseline \cap SR \cap \neg MR’ row shows the number of false positives predicted by both the baseline model and the model using SR data augmentation, but not by the one using MR data augmentation.

3.4.2 A closer look at errors

The next question we aim to answer is how data augmentation improves the performance. Put another way, are data augmentation methods guaranteed to fix some particular errors predicted by the baseline model without using data augmentation, and if yes, which types of errors are more likely to get rectified.

To answer this question, we train three models—one baseline model without using data augmentation, two models using synonym replacement and mention replacement, respectively—and then compare the error predictions by these three models.

From Table 3.4, we find *data augmentations are more likely to reduce false positives than false negatives*. In other words, if the baseline model fails to recall some entity mentions, the model trained using data augmentation usually fails to recall them as well. However, data augmentation can fix those mistakenly predicted entity mentions. On one hand, we believe this improvement can be linked to the *over-fitting* problem. That is, the model trained without using data augmentation may overfit some patterns observed in the small training set, and data augmentation can relieve this problem, by creating a new combination of mention

and context. On the other hand, training model using data augmentation provides very little improvement on fixing those false negatives. Note that data augmentation may also make large amount of new false positives. In other words, there is no guarantee data augmentation can fix some particular errors predicted by models without using data augmentation, since they may provide a mechanism to prevent the training from over-fitting, but not help the learning algorithm to discover new regularities.

3.5 Summary

We design several easy to use data augmentation methods for NER: label-wise token replacement, synonym replacement, mention replacement, and shuffle within segments. Through experiments on two datasets from the biomedical domain, we find that all proposed data augmentation methods can improve over the strong baseline, where large scale pre-trained models are used, and synonym replacement outperforms other augmentation on average.

Cost-effective Selection of Pre-training Data

Pre-training language representation models on unlabelled data and then adapting them to downstream supervised tasks has become a standard practice in NLP. However, the selection of pre-training data usually resorts to intuition, which varies across NLP practitioners. We make use of similarity measures to nominate in-domain pre-training data. Experimental results suggest that simple similarity measures are good predictors of the usefulness of pre-trained language representation models on downstream NER tasks.

4.1 Overview

Sequential transfer learning—which pre-trains a model from a *source* task and then adapts it to a different *target* task—has demonstrated its effectiveness on a range of NLP tasks (Pan and Yang, 2009; Weiss et al., 2016; Ruder, 2019). There are two stages in this procedure: pre-training, and adaptation. Researchers who work on low-resource NLP usually spend a considerable amount of efforts and resources on choosing useful external data sources and investigating how to transfer knowledge to their target tasks.

Mikolov et al. (2013b); Peters et al. (2018); Devlin et al. (2019) make the most of limited labelled data by incorporating language representation models which are pre-trained on a large amount of unlabelled data. This benefits a range of NLP tasks where appropriate unlabelled data is available, and has become a standard practice in NLP.

However, there is still a lack of systematic study on how to select appropriate data to pre-train language representation models. We observe two heuristic strategies in the literature:

- (1) collecting as large as possible generic data, such as news (Mikolov et al., 2013b; Peters et al., 2018; Liu et al., 2019), web crawl (Pennington et al., 2014; Mikolov et al., 2018), and Wikipedia (Bojanowski et al., 2017; Devlin et al., 2019); and,
- (2) selecting moderate size data focusing on a specific domain. The resulting pre-trained models are called *domain-specific models* (Chiu et al., 2016; Karimi et al., 2017; Chronopoulou et al., 2019; Nguyen et al., 2020; Lee et al., 2020).

The advantage of the first strategy is that the pre-trained generic models can be re-used in various domains, however, the corresponding training cost is high and unbearable to many academic labs. For example, Liu et al. (2019) pre-train the RoBERTa model using 1024 V100 GPUs, which are only accessible by large companies. Therefore, we focus on studying the second strategy, and we aim to pre-train domain-specific models, optimising the performance on downstream biomedical NER datasets.

Studies on domain-specific language representation models empirically show that target task performance can be improved, when in-domain data is used for pre-training (Alsentzer et al., 2019; Lee et al., 2020; Beltagy et al., 2019). These publicly available domain-specific models are valuable to the NLP community. However, the selection of in-domain data usually resorts to intuition, which varies across NLP practitioners (Section 4.2). According to Halliday and Hasan (1989), the context specific usage of language is affected by three factors: *field* (the subject matter being discussed), *tenor* (the relationship between the participants in the discourse and their purpose) and *mode* (communication medium, such as ‘spoken’ or ‘written’). Generally, the selection of pre-training data in existing domain-specific models is mainly based on the field rather than the tenor. For example, BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019) are both pre-trained on scholar articles, but on different fields (biology and computer science).

We first show in Section 4.2 that human intuition regarding selecting pre-training data varies across practitioners, motivating our work on employing quantitative measures to nominate in-domain pre-training data. We then describe several measures which can quantify similarity between two datasets in Section 4.3. We pre-train several domain-specific language representation models on different sources, and investigate their effectiveness on various

downstream NER datasets, respectively (Section 4.4). Finally, through correlation analysis, we show that simple similarity measures can be used to nominate in-domain pre-training data (Section 4.5.1).

4.2 What Human Intuition Indicates

We surveyed 30 NLP or machine learning practitioners to learn the human intuition regarding selection of pre-training data. Participants were provided short descriptions of the target data T , and two possible source data $S1$ and $S2$ as

- T : Online forum posts about medications;
- $S1$: Research papers about biology and health;
- $S2$: Online reviews about restaurants, hotels, barbers, mechanics, etc.

A screenshot is shown in Figure 4.1.

We constructed each of the descriptions as ‘ t about f ’ where t is intended to indicate the tenor and f the field. Each participant rated both sources on a five-point Likert, indicating agreement with the statement “*Unsupervised pre-training on S would be useful for supervised named entity recognition learning on T* ”.

Survey results show that 73% of the participants agreed or strongly agreed that $S1$ —sharing similar *biomedical* field with the target—would be useful. Only 27% agreed that $S2$ —sharing similar *social media* tenor with the target—would be useful (Figure 4.2). On the one hand, a Wilcoxon signed-rank test indicates that scores are significantly higher for $S1$ than for $S2$ ($Z = 43.0, p < 0.001$). On the other hand, these results show the variety across practitioners, motivating our work on employing quantitative measures to nominate pre-training data. Our empirical investigations (detailed in Section 4.4) also suggest that human intuition maybe unreliable regarding selecting pre-training data. That is, practitioners favour field over tenor when selecting pre-training data, and this would be detrimental to accuracy of the target NER tasks.

Survey

Hi,

We are conducting a research to find a cost-effective method to select unsupervised pretraining data for NER (Named Entity Recognition).

We are hoping you can help us to get a rough understanding about human judgement on this question.

This survey may take 1 minute.

Corpora description

T: Online forum posts about medications.

S1: Research papers about biology and health.

S2: Online reviews about restaurants, hotels, barbers, mechanics, etc.

The size of S1 and S2 is similar, both of which is much larger than the size of T.

Questions

1. Do you think unsupervised pretraining models on S1 would be useful for supervised named entity learning on T?

Mark only one oval.

- strongly agree
 agree
 neutral
 disagree
 strongly disagree

2. Do you think unsupervised pretraining models on S2 would be useful for supervised named entity learning on T?

Mark only one oval.

- strongly agree
 agree
 neutral
 disagree
 strongly disagree

3. Your background (Please select the first suitable option from top to bottom)

Mark only one oval.

- I work on Biomedical NLP
 I work on NLP
 I work on Machine Learning
 I work on Compute Science
 Other: _____

This content is neither created nor endorsed by Google.

Google Forms

FIGURE 4.1. Survey questions regarding selection of pre-training data.

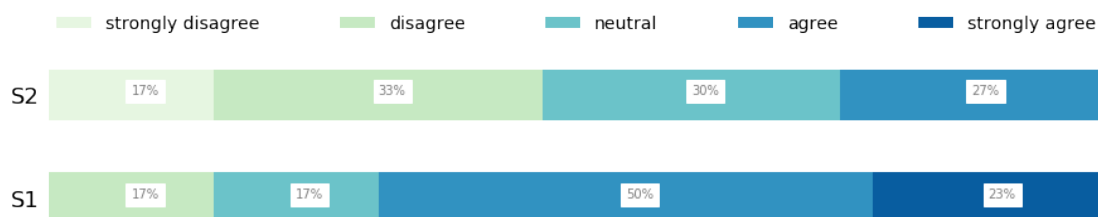


FIGURE 4.2. Likert scale ratings from NLP and ML practitioners ($N = 30$) for the statement ‘*Unsupervised pre-training on S would be useful for supervised named entity recognition learning on T.*’ Target data T is described as ‘*Online forum posts about medications,*’ source data S1 as ‘*Research papers about biology and health,*’ and source data S2 as ‘*Online reviews about restaurants, hotels, barbers, mechanics, etc.*’

4.3 Similarity Measures

Recall that the context specific usage of language is affected by three factors: field, tenor and mode (Halliday and Hasan, 1989). Researchers who select pre-training data from a similar field believe that, if the source data has a similar field to the target data, they tend to share similar topical vocabulary. Conversely, vocabularies are different from each other if source and target are from different fields. Imagine datasets about medications and restaurants. Those who select pre-training data from a similar tenor believe that tenor may impact the writing style of text. Imagine the participants in online reviews and scientific papers, their relationships to each other, their purposes and how these affect text style, including punctuation, lexical normalisation, politeness, emotiveness and so on (Lee, 2001; Solano-Flores, 2006). We do not explicitly consider mode, because all of the datasets studied in this thesis are written text.

Below, we detail different measures based on these intuitions to quantify different aspects of similarity between two datasets.

4.3.1 Target vocabulary covered

The first measure is simply the percentage of the target vocabulary that is also present in the source data. An extremely dissimilar example is that of different languages. They have a totally different vocabulary and are considered dissimilar, even if they are written in a

similar style and talking about the same subject. Note that our focus is on transferring through pre-trained models using one single source and we do not consider multilingual similarity. We propose *Target Vocabulary Covered (TVC)* as a measure of field, calculated as

$$TVC(D_S, D_T) = \frac{|V_{D_S} \cap V_{D_T}|}{|V_{D_T}|},$$

where V_{D_S} and V_{D_T} are sets of unique content words (nouns, verbs, adjectives) in source and target datasets respectively.

4.3.2 Jaccard similarity of vocabularies

Jaccard similarity coefficient (Agresti, 2003), is a statistic used for estimating the similarity and diversity of two sets. By calculating

$$JSC(D_S, D_T) = \frac{|V_{D_S} \cap V_{D_T}|}{|V_{D_S} \cup V_{D_T}|},$$

Jaccard Similarity of Vocabularies (JSV) can be used to measure the similarity between source and target vocabularies, meanwhile, factoring out the source vocabulary size.

4.3.3 Language model perplexity

A language model (Schütze et al., 2008) assigns a probability to any sequence of words $[w_1, \dots, w_N]$ using chain rule of probability:

$$p(w_1, w_2, \dots, w_N) = \prod_{i=1}^N p(w_i | w_1^{i-1}),$$

where N is the length of the sequence and w_1^{i-1} are all words before word w_i . In practice, this equation can be simplified by n-gram models based on Markov Assumption:

$$p(w_1, w_2, \dots, w_N) = \prod_{i=1}^N p(w_i | w_{i-n+1}^{i-1}),$$

where w_{i-n+1}^{i-1} represents only n preceding words of w_i . To make the model generalise better, smoothing techniques can be used to assign non-zero probabilities to unseen events. We use

Kneser-Ney smoothed 3-gram models (Heafield, 2011) to measure the similarity between two datasets. Specifically, we first train the language model on the source data, then evaluate it on the target data using perplexity to represent the degree of similarity. The intuition is that, if the model finds a sentence very unlikely (dissimilar from the data where this language model is trained on), it will assign a low probability and therefore high perplexity. The summed up *perplexity (PPL)* is then:

$$PPL(D_S, D_T) = \sum_{i=1}^m P(D_T^i)^{-\frac{1}{N_i}},$$

where m is the number of sentences in the target data set, and $P(D_T^i)$ is the probability assigned by the language model trained on the source data to the i -th sentence from the target data set, whose sentence length is N_i .

Similar to TVC, PPL is token-based but also captures surface structure. We therefore propose PPL as a proxy to measure tenor as well as field.

4.3.4 Jensen-Shannon divergence

Jensen-Shannon divergence (JSD), based on term distributions, has been successfully used for domain adaptation (Ruder and Plank, 2017). We first measure the probability of each term (up to 3-gram) in source S and target data T , separately. Then, we use the Jensen-Shannon divergence (Fuglede and Topsoe, 2004) between these two probability distributions

$$JSD(S||T) = \frac{1}{2}KL(S||M) + \frac{1}{2}KL(T||M),$$

where

$$KL(P||Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

and

$$M = \frac{1}{2}(S + T)$$

as a proxy to measure tenor as well as field.

4.4 Datasets

We use six datasets as source data, covering a range of fields and tenors:

News: The original one billion word language model benchmark data (Chelba et al., 2013), produced from News Crawl data. *Popular reporting* usually involves one or several writers, and a large number of readers. The text is usually edited several times for easy understanding.

Books: A corpus of 11,038 books in 16 different genres, e.g., Romance, Fantasy, Science fiction, etc., collected by Zhu et al. (2015). These books are all free books written by yet unpublished authors. *Fiction books* usually involves one writer, and a moderate size of readers. The text is usually edited many times, reflecting writer personality.

MIMIC: A clinical database comprising over 58,000 hospital admissions for intensive care unit (ICU) patients (Johnson et al., 2016). *Clinical notes* are usually written by doctors and nurses under time pressure, and read by their colleagues. The text is seldom carefully edited, so it is syntactically noisy and usually contains a lot of jargon for efficient communication.

PubMed: Titles and abstracts of biomedical scholar articles. *Scholar articles* are usually written by a small groups of writers, and the readers usually have similar knowledge background with authors. The text is usually edited many times for more comprehensible and less ambiguous.

Yelp: Crowd-sourced reviews about local businesses, including restaurants, hotels, barbers, mechanics, etc. *Online review* is usually written by a single customer, and read by a small group of people who are interested in the business. The text is usually edited once, and writer tends to use descriptive language to share their experiences.

Wikipedia: A free online encyclopedia. *Online encyclopedia* is created and edited by volunteers around the world, and it has around 250 million page views every day ¹. Another important feature of Wikipedia as an open-collaborative website is that articles are edited all the time by human users and bots, reflecting the newest development of the world knowledge.

¹Siteviews Analysis: shorturl.at/ayQR5. Accessed data: 2021-Jan-31.

Source	# sentences	# tokens	# unique tokens	Avg. sentence length
Books	53.9M	0.69B	0.5M	12.8
MIMIC	61.9M	0.61B	0.5M	9.8
News	27.5M	0.70B	2.1M	25.4
PubMed	29.3M	0.69B	4.2M	23.5
Wikipedia	31.1M	0.69B	3.3M	22.1
Yelp	50.7M	0.69B	1.0M	13.5

TABLE 4.1. Descriptive statistics of the source datasets.

To isolate the impact of source data size, we randomly sample all source data to approximately 700 million tokens. The only exception is on MIMIC. Although all text from MIMIC data set has been used, it is still relatively small, comparing to other sources. The data statistics of source data is listed in Table 4.1. Based on the number of unique tokens and average sentence length, we can see these sources are roughly split into two categories: formal text—PubMed, Wikipedia and News—with large vocabulary and long sentences, and informal text—Books, MIMIC, and Yelp—with small vocabulary and short sentences.

Ten NER datasets are used as target data: BC2GM (BioCreative II Gene Mention Recognition) (Smith et al., 2008), BTC (Broad Twitter Corpus) (Derczynski et al., 2016), CADEC (CSIRO Adverse Drug Event Corpus) (Karimi et al., 2015a), CoNLL 2003 (Sang and Meulder, 2003), EBM (Evidence Based Medicine) (Nye et al., 2018), i2b2 2010 (Uzuner et al., 2011), JNLPBA (Kim et al., 2004), NCBI-DISEASE (Doğan et al., 2014), SciERC (Luan et al., 2018), WetLab (Kulkarni et al., 2018), and W-NUT 2016 (Strauss et al., 2016). Details of these target data are listed in Table 4.2.

4.5 Experimental Results

Similarity Between Source and Target Datasets. The results shown in Table 4.3 and 4.4 indicate that PubMed is the most similar source to most of these target datasets from the Biomedical domain. It achieves lower language model perplexity, higher target vocabulary covered, and Jensen-Shannon Divergence when evaluated against BC2GM, EBM, JNLPBA, NCBI-disease, SciERC and Wetlab compared to other sources. On one hand, it is expected

Target	Entity Categories	Description
BC2GM	Gene	Biomedical scholar articles
BTC	Person, Organisation, Location	Tweets sampled across different regions, temporal periods, and types of Twitter users
CADEC	Adverse Drug Event, Disease, Drug, Finding, Symptom	Posts taken from AskaPatient, which is a forum where consumers can discuss their experiences with medications.
EBM	Intervention, Outcome and Comparator	Scholar articles about clinical trials
i2b2 2010	Problem, Treatment and Test	Clinical notes about health
JNLPBA	Protein, DNA, RNA, Cell line and Cell type	Abstract of journal articles about biology.
NCBI-disease	Disease	Abstract of journal articles about health.
SciERC	Generic, Material, Method, Metric, Other-Scientific-Term, Task	Journal articles about Computer Science, Material Sciences and Physics
Wetlab	Action, 9 object-based (Amount, Concentration, Device, Location, Method, Reagent, Speed, Temperature, Time) entity types, 5 measure-based (Numerical, Generic-Measure, Size, pH, Measure-Type) and 3 other (Mention, Modifier, Seal) types	Protocols written by researchers about conducting biology and chemistry experiments.

TABLE 4.2. List of the target NER datasets and their specifications.

that PubMed is similar to BC2GM, EBM, JNLPBA, NCBI-disease and Wetlab, since they are all scientific writing about biology and health, thus being similar in terms of both field and tenor. On the other hand, although SciERC does not have the same field as PubMed (computer science, material and physics versus biology and health), they are similar because they share a similar tenor (scholarly publications). On i2b2-2010 (clinical notes), only the target vocabulary covered measure indicates PubMed is the most similar source, whereas other three metrics indicate MIMIC as the most similar source.

The second observation is that tenor might be reflected more than field by these measures. Source data Yelp is more similar to CADEC than PubMed and MIMIC from both language model perplexity and Jensen-Shannon Divergence perspectives. CADEC is a data set focusing on recognising drugs, diseases and adverse drug events. The field of CADEC is therefore

Target	Source	Similarity			
		TVC (%)	JSV (%)	PPL	JSD
BC2GM	Books	37.39	9.54	109.65	36.29
	MIMIC	41.95	13.69	101.39	38.17
	News	48.12	6.15	101.34	38.02
	PubMed	81.20	7.24	75.43	48.32
	Wikipedia	60.19	5.05	93.10	39.28
	Yelp	36.26	8.70	109.11	37.01
BTC	Books	47.96	10.09	61.63	38.44
	MIMIC	26.80	6.83	68.54	33.92
	News	54.92	5.58	59.49	36.61
	PubMed	41.97	2.84	71.44	33.48
	Wikipedia	54.46	3.57	61.19	35.00
	Yelp	47.58	9.40	60.58	39.22
CADEC	Books	80.16	5.08	47.37	42.72
	MIMIC	78.21	6.54	47.92	38.21
	News	85.59	2.47	45.90	39.67
	PubMed	82.03	1.57	52.38	37.69
	Wikipedia	84.41	1.55	49.41	38.18
	Yelp	81.89	4.85	45.52	44.82
EBM	Books	29.68	10.29	146.01	36.03
	MIMIC	32.61	14.06	130.42	37.70
	News	44.30	8.21	125.28	38.47
	PubMed	70.66	9.30	91.87	51.85
	Wikipedia	47.20	5.79	125.61	39.00
	Yelp	29.94	9.85	142.15	36.82

TABLE 4.3. Similarity values measured between source and target datasets. TVC: Target Vocabulary Covered. JSV: Jaccard similarity of Vocabularies. PPL: language model perplexity. JSD: Jensen-Shannon Divergence based on term distributions.

more similar to PubMed which includes journal articles in health discipline and MIMIC which contains clinical notes. However, CADEC is written by patients, and can be considered as ‘drug reviews’. The tenor is therefore closer to the one in Yelp, where customers use informal language to describe their experiences. Target vocabulary covered nominates News as the most similar source. Note that News has a moderate size vocabulary, 2.1 millions unique tokens, whereas the vocabulary size of PubMed is 4.2 million.

Target	Source	Similarity			
		TVC (%)	JSV (%)	PPL	JSD
i2b2 2010	Books	44.65	5.74	45.55	37.75
	MIMIC	58.36	9.95	29.29	48.99
	News	56.28	3.42	43.32	37.30
	PubMed	64.86	2.64	38.92	38.23
	Wikipedia	59.59	2.32	42.24	37.94
	Yelp	45.76	5.52	44.75	37.92
JNLPBA	Books	31.25	5.87	105.66	35.84
	MIMIC	32.30	7.74	101.82	36.25
	News	40.46	3.72	97.64	37.31
	PubMed	71.49	4.52	65.00	47.80
	Wikipedia	46.46	2.78	91.20	38.29
	Yelp	30.46	5.38	107.22	36.55
NCBI-disease	Books	54.78	5.05	95.03	36.07
	MIMIC	57.02	6.90	90.02	37.24
	News	65.84	2.82	86.97	37.43
	PubMed	87.15	2.50	64.48	44.46
	Wikipedia	76.17	2.08	80.70	38.39
	Yelp	52.78	4.55	95.63	36.82
SciERC	Books	70.19	4.34	88.42	36.11
	MIMIC	57.06	4.60	93.64	35.31
	News	77.67	2.19	80.88	37.03
	PubMed	84.14	1.58	71.03	39.72
	Wikipedia	81.23	1.46	77.72	37.56
	Yelp	67.37	3.88	89.39	36.89
Wetlab	Books	48.82	3.97	60.62	36.43
	MIMIC	44.69	4.74	58.02	36.52
	News	58.06	2.19	57.47	36.11
	PubMed	68.54	1.72	52.11	37.15
	Wikipedia	61.24	1.47	55.99	36.41
	Yelp	50.29	3.83	57.84	37.12

TABLE 4.4. Similarity values measured between source and target datasets (continued). TVC: Target Vocabulary Covered. JSV: Jaccard similarity of Vocabularies. PPL: language model perplexity. JSD: Jensen-Shannon Divergence based on term distributions.

The last observation is that using different measures can lead to almost the same answer regarding *which source is the most similar one to a given target*, except for the Jaccard similarity of vocabularies. Using Jaccard similarity of vocabularies measure, MIMIC source

is nominated as the most similar source against target sets, except for BTC. This might be explained by the fact that the vocabulary size of MIMIC is the smallest one in all sources, and Jaccard similarity of vocabularies measure favours sources with small vocabulary size than the ones with large vocabulary size.

Effectiveness of domain-specific models on downstream NER tasks. After we quantify the similarity between source and target datasets, the next step is to investigate the impact of source data on pre-trained language representation models. We pre-train ELECTRA (Clark et al., 2020)—a sample-efficient variant of BERT—on different sources separately, then observe how the effectiveness of these pre-trained models varies in different downstream NER datasets.

The most common approach of training domain-specific models is *continue pre-training*, which is used by BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), and BERTtweet (Nguyen et al., 2020). Continue pre-training approach starts from an existing pre-trained model—usually pre-trained on large size generic data set—and continues training on a domain-specific corpus. The main advantage of continue pre-training is that they can inherit knowledge from language representation models pre-trained on generic data, and thus be considered as capturing both generic domain and domain-specific knowledge. However, we aim to investigate the impact of pre-training data, therefore, we use the *learning from scratch* approach, eschewing the potential impact of generic data. We follow the hyper-parameter setting in (Clark et al., 2020), shown in Table 4.5 to train the domain-specific models. Training of each model took four days using 1 Nvidia Tesla v100 GPU.

Evaluation results using these domain-specific models on downstream NER tasks show that the effectiveness varies in different target datasets (Table 4.6). In other words, no single source is suitable for all target NER datasets. It is worthy note that most of these results (for example on BC2GM, BTC, CADEC, SciERC) are lower than state-of-the-art results on these datasets with large margin. This is mainly because our pre-trained models are smaller—smaller hidden size, less number of attention heads, smaller embedding size—than the ones in other studies.

Hyper-parameter	Value
Number of layers	12
Hidden size	256
Intermediate size	1024
Attention heads	4
Attention head size	64
Embedding size	128
Learning rate	5e-4
Train steps	800K
Vocab size	30,994

TABLE 4.5. Pre-train hyper-parameters, which follow the practice of training ELECTRA-SMALL in (Clark et al., 2020).

	Book	MIMIC	News	PubMed	Wiki	Yelp
BC2GM	72.4 (0.2)	72.7 (0.5)	74.4 (0.4)	80.7 (0.2)	75.1 (0.4)	72.4 (0.3)
BTC	70.4 (0.2)	63.1 (0.3)	75.2 (1.3)	67.1 (0.7)	74.9 (0.4)	70.8 (0.5)
CADEC	64.3 (0.6)	67.5 (0.3)	65.1 (0.3)	66.1 (0.5)	65.5 (0.5)	66.1 (0.6)
EBM	39.8 (0.4)	41.0 (0.4)	41.1 (0.6)	43.5 (0.3)	40.7 (0.4)	40.8 (0.5)
i2b2-2010	78.7 (0.3)	87.7 (0.2)	79.8 (0.9)	85.1 (0.3)	79.6 (0.4)	79.4 (0.3)
JNLPBA	69.7 (0.2)	70.1 (0.3)	70.1 (0.7)	73.1 (0.2)	70.7 (0.2)	70.0 (0.2)
NCBI-Disease	77.7 (0.5)	80.3 (0.7)	77.9 (4.8)	85.8 (0.5)	80.8 (0.3)	79.2 (0.8)
SciERC	37.2 (4.1)	23.8 (1.3)	25.9 (2.2)	41.0 (20.6)	47.3 (1.4)	38.7 (3.1)
WetLab	78.2 (0.1)	78.1 (0.2)	78.1 (0.3)	78.7 (0.2)	78.3 (0.1)	78.0 (0.1)

TABLE 4.6. The effectiveness of domain-specific pre-trained models on downstream NER tasks. We report the mention level F_1 scores.

The best performing models on each target data are all pre-trained on the most similar source which is nominated by at least one similarity measure, except for SciERC (Table 4.3 and 4.4).

4.5.1 Predictiveness of similarity measures

To analyse how proposed similarity measures can be used to nominate the best pre-training data option, we investigate the correlation between these similarity values and the effectiveness of pre-trained models on target tasks. Specifically, we employ Spearman rank-order correlation coefficient to measure the relationship between the ranking of similarity values and NER results. For example, given the target data set NCBI-disease, the rank of sources is PubMed,

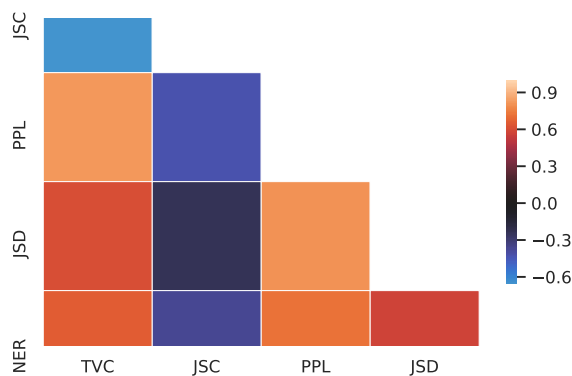


FIGURE 4.3. Correlation between different similarity measures and the effectiveness of domain-specific pre-trained models.

Wikipedia, MIMIC, Yelp, News, and Book, if they are sorted based on the effectiveness of different domain-specific models. Similarity, if they are sorted based on the target vocabulary covered measure, the rank of sources is PubMed, Wikipedia, News, MIMIC, Books, and Yelp. The Spearman rank-order correlation coefficient between these two rankings is 0.71.

The results in Figure 4.3 show that these proposed similarity measures are predictive of the effectiveness of the pre-training data, except for Jaccard similarity of vocabularies.

4.5.2 Comparison to publicly available pre-trained models

Literature shows substantial improvements are sometimes possible when pre-training on large generic corpora (Liu et al., 2019; Baevski et al., 2019). Given that pre-trained models are freely available, is it even necessary to pre-train on similar data as proposed above? We compare to publicly available ELECTRA models trained on 3.3 Billion tokens generic data. Note that the publicly available model we choose has the same model size. It is also pre-trained using the same hyper-parameters as ours, except that it is pre-trained longer than ours (1.45M vs.0.8M).

	Domain-specific model	Generic domain model
BC2GM	80.7 (0.2)	63.6 (0.7)
BTC	75.2 (1.3)	75.3 (0.2)
CADEC	67.5 (0.3)	57.9 (0.3)
EBM	43.5 (0.3)	41.6 (0.2)
i2b2-2010	87.7 (0.2)	82.5 (0.1)
JNLPBA	73.1 (0.2)	71.2 (0.4)
NCBI-Disease	85.8 (0.5)	81.8 (0.6)
SciERC	47.3 (1.4)	47.3 (0.5)
WetLab	78.7 (0.2)	78.8 (0.1)

TABLE 4.7. Comparison between our best performing domain-specific models and the publicly available generic domain model.

These results, shown in Table 4.7 indicate that a small similar source reduces the computational cost without sacrificing the performance. This is especially important in practice, because collecting data and pre-training models are expensive, in terms of both computational and environmental cost (Schwartz et al., 2019).

4.6 Summary

This chapter focuses on whether there are cost-effective methods to nominate datasets to pretrain language representation models that are building blocks of NER models. We propose using different measures to measure different aspects of similarity between source and target data. We investigate how these measures correlate with the effectiveness of pre-trained models for NER tasks. While different NLP tasks may rely on different aspects of language, our study is a step towards systematically guiding researchers on their choice of data for pre-training, and models pre-trained on small size domain-specific corpus can outperform the one pre-trained on large size generic domain data.

Transition-based Model for Discontinuous NER

Discontinuous mentions represent compositional concepts, for example disorders or symptoms, that differ from concepts represented by individual components, for example body locations or general feelings. In downstream applications such as pharmacovigilance and summarization, recognising these discontinuous mentions is more useful than recognising separate components. We propose a transition-based model that can effectively recognise discontinuous mentions without sacrificing the accuracy on continuous mentions.

5.1 Overview

NER is a critical component of biomedical text mining applications. In pharmacovigilance, it can be used to identify adverse drug events in consumer reviews in online medication forums, alerting medication developers, regulators, and clinicians (Leaman et al., 2010; Sarker et al., 2015; Karimi et al., 2015b). In clinical settings, NER can be used to extract and summarise key information from electronic medical records such as conditions hidden in unstructured doctors' notes (Feblowitz et al., 2011; Wang et al., 2018b). These applications require identification of complex entity mentions, discontinuous and overlapping mentions, not seen in generic domains.

Widely used sequence tagging techniques encode two assumptions that do not always hold: (1) mentions do not overlap, therefore each token can belong to at most one mention; and, (2) mentions comprise continuous sequences of tokens. Nested entity recognition addresses violations of the first assumption (more discussions in Section 2.4). However, the violation

The left atrium is mildly dilated .
E1 E1

have much muscle pain and fatigue .
E2 E3

FIGURE 5.1. Examples involving discontinuous mentions, taken from the SHARE/CLEF 13 (Pradhan et al., 2013) and CADEC (Karimi et al., 2015a) datasets, respectively. The first example contains a discontinuous mention ‘*left atrium dilated*’, the second example contains two mentions that overlap: ‘*muscle pain*’ and ‘*muscle fatigue*’ (discontinuous).

of the second assumption is comparatively less studied and requires handling discontinuous mentions (see examples in Figure 5.1).

In contrast to continuous mentions which are often short spans of text, discontinuous mentions consist of *components* that are separated by *intervals*. Recognising discontinuous mentions is particularly challenging as exhaustive enumeration of possible mentions, including discontinuous and overlapping spans, is exponential to sentence length. Existing approaches for discontinuous NER either suffer from high time complexity (McDonald et al., 2005) or ambiguity in translating intermediate representations into mentions (Tang et al., 2013a; Metke-Jimenez and Karimi, 2016; Muis and Lu, 2016). In addition, current arts use traditional approaches that rely on manually designed features, which are tailored to recognise specific entity categories. Also, these features usually do not generalise well in different types of text (Leaman et al., 2015).

Motivations. The main motivation for recognising discontinuous mentions is that they usually represent *compositional concepts* that differ from concepts represented by individual components. For example, the mention ‘*left atrium dilated*’ in the first example of Figure 5.1 describes a disorder which has its own CUI (Concept Unique Identifier) in UMLS (Unified Medical Language System), whereas both ‘*left atrium*’ and ‘*dilated*’ also have their own CUIs. In downstream applications such as pharmacovigilance and summarization, recognising these discontinuous mentions that refer to disorders or symptoms is more useful than recognising separate components which may refer to body locations or general feelings.

Another motivation for discontinuous NER is that discontinuous mentions usually overlap, and separating these overlapping mentions rather than identifying them as a single mention is important for downstream tasks, such as entity linking where the assumption is that the input mention refers to one entity (Shen et al., 2014).

In this chapter, we first characterise three datasets, from the biomedical domain, with a substantial number of discontinuous mentions (Section 5.2). Then, we introduce a transition-based model that can recognise discontinuous mentions (Section 5.3). Through experiments, we show that our model can effectively recognise discontinuous mentions without sacrificing the accuracy on continuous mentions (Section 5.4). Analysis also suggests that our model is better than existing discontinuous NER models at handling long mentions and mentions that do not overlap or overlap at left, resulting in higher recall (Section 5.5).

5.2 Datasets

Although some text annotation tools, such as BRAT (Stenetorp et al., 2012), allow discontinuous annotations, corpora annotated with large number of discontinuous mentions are still rare because they are hard to annotate. We describe three datasets from the biomedical domain that include a substantial number of discontinuous mentions: CADEC (Karimi et al., 2015a), SHARE/CLEF 2013 (Pradhan et al., 2013) and SHARE/CLEF 2014 (Mowery et al., 2014). We then motivate the discontinuous NER task.

CADEC corpus is sourced from posts from *AskaPatient*¹, a forum where patients can discuss their experiences with medications. The entity categories annotated in CADEC include drug, Adverse Drug Event (ADE), disease and symptom. In this work, we only consider the ADE annotations because only the ADEs involve discontinuous mentions. Note that ADEs in CADEC are defined as span of text that are clearly associated with a drug and should have the corresponding MEDDRA (Medical Dictionary for Regulatory Activities) term. SHARE/CLEF 2013, specifically Task 1 of the SHARE/CLEF eHealth evaluation lab 2013, focuses on identification of disorder mentions in clinical reports. The corpus is

¹<https://www.askapatient.com/>. Accessed data: 22nd May 2021

sourced from de-identified clinical reports, including discharge summaries, electrocardiogram, echocardiogram, and radiology reports (Johnson et al., 2016). A disorder mention is defined as any span of text which can be mapped onto a concept in the Disorder semantic group of SNOMED-CT (Systematised Nomenclature of Medicine – Clinical Terms). SHARE/CLEF 2014, an extension of the SHARE/CLEF 2013 task, focuses on template filling of disorder attributes. That is, given a disorder mention and its surrounding words, recognise the attributes of the disorder mention from its context, including subject class, severity indicator, uncertainty indicator. In this work, we frame SHARE/CLEF 2014 as a disorder-NER dataset. That is, we consider only disorder annotations, without taking their attributes into consideration.

Descriptive statistics of these three datasets is listed in Table 5.1. On average, discontinuous mentions are longer than continuous mentions, because they consist of several components, and the intervals between different components make the total length of span even longer. Another important characteristic of discontinuous mentions is that they usually *overlap*. That is, several mentions may share components that refer to the same body location (e.g., ‘*muscle*’ in ‘*muscle pain and fatigue*’), or the same feeling (e.g., ‘*Pain*’ in ‘*Pain in knee and foot*’). From this perspective, we also categorise discontinuous mentions into the following groups:

- No overlap: in such cases, the discontinuous mention can be intervened by severity indicators (e.g., ‘*is mildly*’ in sentence ‘*left atrium is mildly dilated*’), preposition (e.g., ‘*on my*’ in sentence ‘*...rough on my stomach...*’) and so on. This category accounts for half of discontinuous mentions in the SHARE/CLEF datasets but only 12% in CADEC.
- Left overlap: the discontinuous mention shares one component with other mentions, and the shared component is at the beginning of the discontinuous mention. This is usually accompanied with coordination structure (e.g., the shared component ‘*muscle*’ in ‘*muscle pain and fatigue*’). Conjunctions (e.g., ‘*and*’, ‘*or*’) are clear indicators of the coordination structure. However, clinical notes (SHARE/CLEF datasets) are usually written by practitioners under time pressure. They often use commas or slashes rather than conjunctions. This category accounts for more than half of discontinuous mentions in CADEC and one third in SHARE/CLEF.

Dataset			
	CADEC	SHARE/CLEF 13	SHARE/CLEF 14
Text type	online posts	clinical notes	clinical notes
Entity type	ADE	Disorder	Disorder
# Documents	1,250	298	433
# Tokens	121K	264K	494K
# Sentences	7,597	18,767	34,618
# Mentions	6,318	11,161	19,131
# Disc.M	675 (10.6)	1,090 (9.7)	1,710 (8.9)
Avg mention L.	2.7	1.8	1.7
Avg Disc.M L.	3.5	2.6	2.5
Avg interval L.	3.3	3.0	3.2
Discontinuous Mentions			
2 components	650 (95.7)	1,026 (94.3)	1,574 (95.3)
3 components	27 (3.9)	62 (5.6)	76 (4.6)
4 components	2 (0.2)	0 (0.0)	0 (0.0)
No overlap	82 (12.0)	582 (53.4)	820 (49.6)
Overlap at left	351 (51.6)	376 (34.5)	616 (37.3)
Overlap at right	152 (22.3)	102 (9.3)	170 (10.3)
Multiple overlaps	94 (13.8)	28 (2.5)	44 (2.6)
Continuous Mentions			
Overlap	326 (5.7)	157 (1.5)	228 (1.3)

TABLE 5.1. The descriptive statistics of the datasets. ADE: adverse drug events; Disc.M: discontinuous mentions; Disc.M L.: discontinuous mention length, where intervals are not counted. Numbers in parentheses are the percentage of each category. Note that due to sentence segmentation issue, there are 13 and 64 mentions crossing multiple sentences in SHARE/CLEF 2013 and SHARE/CLEF 2014, respectively. We remove these mentions, as we frame the task as a sentence-level NER problem.

- Right overlap: similar to left overlap, although the shared component is at the end. For example, ‘hip/leg/foot pain’ contains three mentions that share the token ‘pain’.
- Multi-overlap: the discontinuous mention shares multiple components with the others, which usually forms *crossing compositions*. For example, the sentence ‘Joint and Muscle Pain / Stiffness’ contains four mentions: ‘Joint Pain’, ‘Joint Stiffness’, ‘Muscle Stiffness’ and ‘Muscle Pain’, where each discontinuous mention share two components with the others.

Although these three datasets – CADEC, SHARE/CLEF 2013 and SHARE/CLEF 2014 – share similar field (the subject matter of the content being discussed), the tenor (the participants in the discourse, their relationships to each other, and their purposes) of CADEC is very different from the SHARE/CLEF datasets. Specially, laymen authors (CADEC) tend to use idioms or ungrammatical phrases to describe their feelings, whereas professional practitioners (SHARE/CLEF) tend to use compact terms for efficient communications. This difference of tenor results in different features of mentions between these datasets. That is, the mentions in CADEC are overall longer than those in SHARE/CLEF datasets, and larger ratio of discontinuous mentions in CADEC are involved in overlapping structure (Table 5.1).

5.3 Proposed Model

We propose a transition-based model based on the shift-reduce parser (Watanabe and Sumita, 2015; Lample et al., 2016) that employs a *stack* to store partially processed spans and a *buffer* to store unprocessed tokens. The learning problem is then framed as: given the state of the parser, predict an action which is applied to change the state of the parser. This process is repeated until the parser reaches the end state, which is the stack and buffer are both empty.

Similar to prior work (Metke-Jimenez and Karimi, 2016; Muis and Lu, 2016) that first predict an intermediate representation of mentions, which are then decoded into the final mentions, our proposed transition-based model uses a sequence of actions as the intermediate representation (refer to Section 2.4).

The main difference between our model and the ones in (Watanabe and Sumita, 2015; Lample et al., 2016) is the set of transition actions. Watanabe and Sumita (2015) use SHIFT, REDUCE, UNARY, FINISH, and IDEA for the constituent parsing system. Lample et al. (2016) use SHIFT, REDUCE, OUT for the flat NER system. Inspired by these models, we design a set of actions specifically for recognising discontinuous and overlapping structure. There is a total of six actions in our model:

- SHIFT moves the first token from the buffer to the stack; it implies this token is part of an entity mention.
- OUT pops the first token of the buffer, indicating it does not belong to any mention.
- COMPLETE pops the top span of the stack, outputting it as an entity mention. If we are interested in multiple entity categories, we can extend this action to COMPLETE- y which labels the mention with entity category y .
- REDUCE pops the top two spans s_0 and s_1 from the stack and concatenates them as a new span which is then pushed back to the stack.
- LEFT-REDUCE is similar to the REDUCE action, except that the span s_1 is kept in the stack. This action indicates the span s_1 is involved in multiple mentions. In other words, several mentions share s_1 which could be a single token or several tokens.
- RIGHT-REDUCE is the same as LEFT-REDUCE, except that s_0 is kept in the stack.

Figure 5.2 shows an example of how the proposed parser recognises entity mentions from a sentence.

5.3.1 Representation of the parser state

Given a sequence of N tokens, we first run a bi-directional LSTM (Graves et al., 2013) to derive the contextual representation of each token. Specifically, for the i -th token in the sequence, its representation can be denoted as:

$$\tilde{c}_i = \left[\overrightarrow{\text{LSTM}}(t_0, \dots, t_i); \overleftarrow{\text{LSTM}}(t_i, \dots, t_{N-1}) \right],$$

where t_i is the concatenation of the embeddings for the i -th token, its character level representation learned using a CNN network (Ma and Hovy, 2016). Pretrained contextual word representations have shown its usefulness on improving various NLP tasks. Here, we can also concatenate pretrained contextual word representations using ELMo (Peters et al., 2018) with \tilde{c}_i , resulting in:

$$c_i = [\tilde{c}_i; \text{ELMO}_i], \quad (5.1)$$

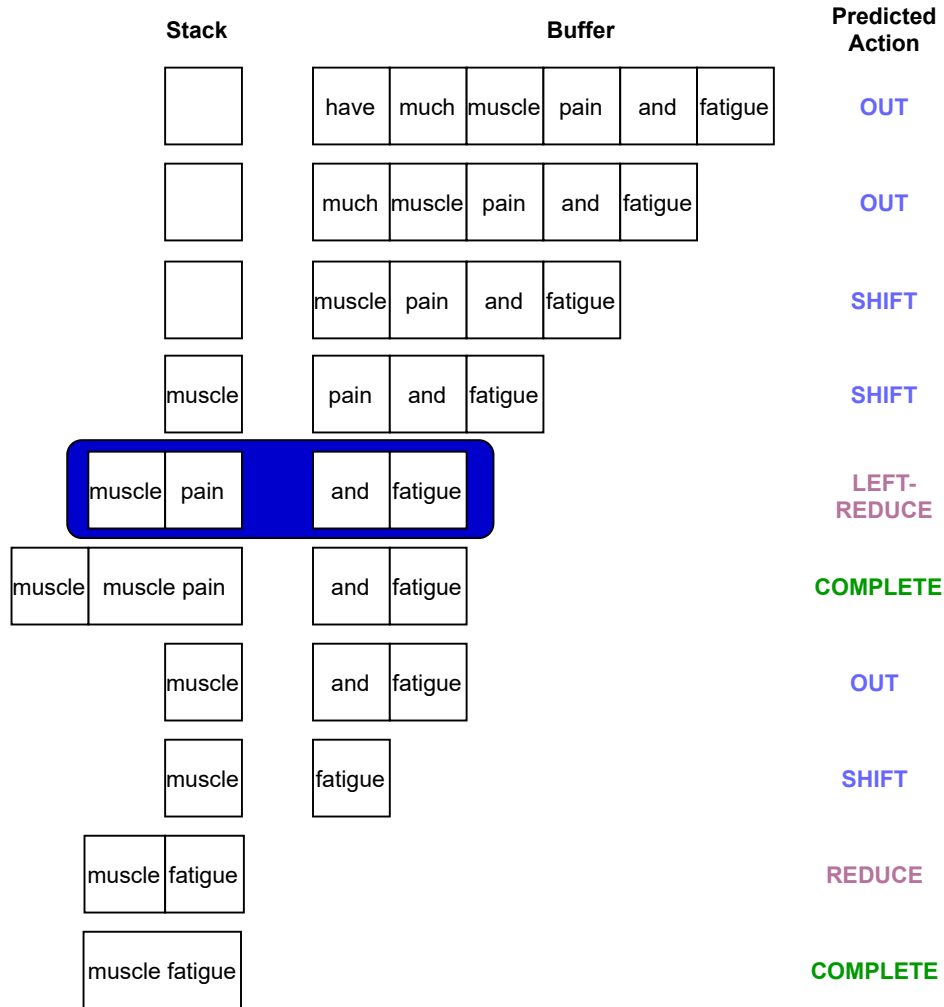


FIGURE 5.2. An example sequence of transitions. Given the states of stack and buffer (blue highlighted), as well as the previous actions, predict the next action (i.e., LEFT-REDUCE) which is then applied to change the states of stack and buffer.

where $ELMO_i$ is the output representation of pretrained ELMO models (frozen) for the i -th token. These token representations c are directly used to represent tokens in the buffer.

Following the work in (Dyer et al., 2015), we use STACKLSTM to represent spans in the stack. That is, if a token is moved from the buffer to the stack, its representation is learned using:

$$s_0 = \text{STACKLSTM}(s_D \dots s_1; c_{\text{SHIFT}}),$$

where D is the number of spans in the stack. Once REDUCE related actions are applied, we use a multi-layer perceptron to learn the representation of the concatenated span. For example, the REDUCE action takes the representation of the top two spans in the stack: s_0 and s_1 , and produces a new span representation:

$$\tilde{s} = \mathbf{W}^T [s_0; s_1] + \mathbf{b}, \quad (5.2)$$

where \mathbf{W} and \mathbf{b} denote the parameters for the composition function. The new span representation \tilde{s} is pushed back to the stack to replace the original two spans: s_0 and s_1 .

5.3.2 Capturing discontinuous dependencies

We hypothesise that the interactions between spans in the stack and tokens in the buffer are important factors in recognising discontinuous mentions. Considering the example in Figure 5.2, a span in the stack (e.g., ‘*muscle*’) may need to combine with a future token in the buffer (e.g., ‘*fatigue*’). To capture this interaction, we use multiplicative attention (Luong et al., 2015) to let the span in the stack s_i learn which token in the buffer to attend, and thus a weighted sum of the representation of tokens in the buffer \mathbf{B} :

$$\begin{aligned} s_i^a &= \text{ATTENTION}(s_i, \mathbf{B}, \mathbf{B}) \\ &= \text{SOFTMAX}(s_i^T \mathbf{W}_i^a \mathbf{B}) \mathbf{B}. \end{aligned} \quad (5.3)$$

We use distinct \mathbf{W}_i^a for spans in different positions s_i separately.

5.3.3 Selecting an action

Finally, we build the parser representation as the concatenation of the representation of top three spans from the stack (s_0, s_1, s_2) and its attended representation (s_0^a, s_1^a, s_2^a), as well as the representation of the previous action \mathbf{a} , which is learned using a simple unidirectional LSTM. If there are less than 3 spans in the stack or no previous action, we use randomly initialised vectors s_{empty} or \mathbf{a}_{empty} to replace the corresponding vector. This parser representation is used as input for the final softmax prediction layer to select the next action.

Note that, given one parser state, not all types of actions are valid. For example, if the stack does not contain any span, only SHIFT and OUT actions are valid because all other actions involve popping spans from the stack. We employ hard constraints that we only select the most likely action from valid actions.

5.4 Experimental Results

To evaluate the effectiveness of our proposed model, we run experiments on previously described three datasets: CADEC, SHARE/CLEF 2013 and SHARE/CLEF 2014, and compare the effectiveness of our model against several baselines.

5.4.1 Baseline models

We choose one flat NER model which is strong at recognising continuous mentions, and two discontinuous NER models as our baseline models:

Flat model. To train the flat model on our datasets, we use an off-the-shelf framework: FLAIR (Akbik et al., 2018), which achieves the state-of-the-art performance on CONLL 2003 dataset. Recall that the flat model cannot be directly applied to datasets containing discontinuous mentions. Following the practice in (Stanovsky et al., 2017), we replace the discontinuous mention with the shortest span that fully covers it, and merge overlapping mentions into a single mention that covers both. Different from (Stanovsky et al., 2017), we apply these changes only on the training set, and not on the development and the test sets.

BIO extension model. The original implementation in (Metke-Jimenez and Karimi, 2016) used a CRF model with manually designed features. We report their results on CADEC in Table 5.2 and re-implement a BiLSTM-CRF-ELMO model using their tag schema (denoted as ‘BIO extension’ in Table 5.2).

Graph-based model. The original paper of (Muis and Lu, 2016) only reported the evaluation results on sentences which contain at least one discontinuous mention. We use their

implementation to train the model and report evaluation results on the whole test set (denoted as ‘Graph’ in Table 5.2). We argue that it is important to see how a discontinuous NER model works not only on the discontinuous mentions but also on all the mentions, especially since, in real datasets, the ratio of discontinuous mentions cannot be made a priori.

5.4.2 Experimental setup

As CADEC does not have an official train-test split, we follow (Metke-Jimenez and Karimi, 2016) and randomly assign 70% of the posts as the training set, 15% as the development set, and the remaining posts as the test set. The train-test splits of SHARE/CLEF 13 and 14 are both from their corresponding shared task settings, except that we randomly select 10% of documents from each training set as the development set. The original SHARE/CLEF 14 task focuses on template filling of disorder attributes: that is, given a disorder mention, recognise the attribute from its context. In this work, we use its mention annotations and frame the task as a discontinuous NER task. Micro average strict match F_1 score is used to evaluate the effectiveness of the model. The trained model which is most effective on the development set, measured using the F_1 score, is used to evaluate the test set. All experiments are repeated five times using different random seeds and averaged results are reported.

5.4.3 Results

When evaluated on the whole test set, our model outperforms three baseline models, as well as over previous reported results in the literature, in terms of recall and F_1 scores (Table 5.2).

The graph-based model achieves highest precision, but with substantially lower recall, therefore obtaining lowest F_1 scores. In contrast, our model improves recall over flat and BIO extension models as well as previously reported results, without sacrificing precision. This results in more balanced precision and recall. Improved recall is especially encouraging for our motivating pharmacovigilance and medical record summarization applications, where recall is at least as important as precision. Note that most of these previous models are tailored for specific entity categories, and utilise domain-specific resources. For example, (Tang et al.,

Model	CADEC			ShARe 2013			ShARe 2014		
	P	R	F	P	R	F	P	R	F
Metke-Jimenez and Karimi (2016)	64.4	56.5	60.2	–	–	–	–	–	–
Tang et al. (2018)	67.8	64.9	66.3	–	–	–	–	–	–
Tang et al. (2013b)	–	–	–	80.0	70.6	75.0	–	–	–
Flat	65.3	58.5	61.8	78.5	66.6	72.0	76.2	76.7	76.5
BIO extension	68.7	66.1	67.4	77.0	72.9	74.9	74.9	78.5	76.6
Graph	72.1	48.4	58.0	83.9	60.4	70.3	79.1	70.7	74.7
Ours	68.9	69.0	69.0	80.5	75.0	77.7	78.1	81.2	79.6

TABLE 5.2. Evaluation results in terms of precision (P), recall (R) and F_1 score (F).

2013b), the best-performing system participated in the SHARE/CLEF 2013 shared task, utilise several external domain-specific resources, such as METAMAP, CTAKES and UMLS. We avoid these tailored resources in our model. We argue that this makes our model more generic and robust, especially since we apply the hyper-parameters tuned on CADEC directly to SHARE/CLEF datasets and obtain similar improvements with respect to the benchmarks.

Effectiveness on recognising discontinuous mentions. Recall that only 10% of mentions in these three datasets are discontinuous. To evaluate the effectiveness of our proposed model on recognising discontinuous mentions, we follow the evaluation approach in (Muis and Lu, 2016) where we construct a subset of test set where only sentences with at least one discontinuous mention are included (Table 5.3). We also report the evaluation results when only discontinuous mentions are considered (Table 5.4). Note that sentences in the former setting usually contain continuous mentions as well, including those involved in overlapping structure (e.g., ‘*muscle pain*’ in the sentence ‘*muscle pain and fatigue*’). Therefore, the flat model, which cannot predict any discontinuous mentions, still achieves 38% F_1 on average when evaluated on these sentences with at least one discontinuous mention, but fails to recognise discontinuous mentions.

Our model again achieves the highest F_1 and recall in all three datasets under both settings. The comparison between these two evaluation results also shows the necessity of comprehensive evaluation settings. The BIO extension model outperforms the graph-based model in terms of F_1 score on CADEC, when evaluated on sentences with discontinuous mentions.

Model	CADEC			SHARE/CLEF 2013			SHARE/CLEF 2014		
	P	R	F	P	R	F	P	R	F
Flat	50.2	36.7	42.4	43.5	28.1	34.2	41.5	31.9	36.0
BIO extension	63.8	52.0	57.3	51.8	39.5	44.8	37.5	38.4	37.9
Graph	69.5	43.2	53.3	82.3	47.4	60.2	60.0	52.8	56.2
Ours	66.5	64.3	65.4	70.5	56.8	62.9	61.9	64.5	63.1

TABLE 5.3. Evaluation results on sentences that contain at least one discontinuous mention.

Model	CADEC			SHARE/CLEF 2013			SHARE/CLEF 2014		
	P	R	F	P	R	F	P	R	F
Flat	0	0	0	0	0	0	0	0	0
BIO extension	5.8	1.0	1.8	39.7	12.3	18.8	8.8	4.5	6.0
Graph	60.8	14.8	23.9	78.4	36.6	50.0	42.7	39.5	41.1
Ours	41.2	35.1	37.9	78.5	39.4	52.5	56.1	43.8	49.2

TABLE 5.4. Evaluation results on discontinuous mentions only.

However, it achieves only 1.8 F_1 when evaluated on discontinuous mentions only. The main reason is that most of discontinuous mentions in CADEC are involved in overlapping structure (88%, cf. Table 5.1), and the BIO extension model is better than the graph-based model at recognising these continuous mentions. On SHARE/CLEF 2013 and 2014, where the portion of discontinuous mentions involved in overlapping is much less than on CADEC, the graph-based model clearly outperforms BIO extension model in both evaluation settings.

Graph based model again achieves highest precision on all three datasets. It also outperforms BIO extension model on SHARE/CLEF 2013 and 2014 in terms of F_1 score, but not on CADEC. Graph based model employs lots of handcrafted features for clinical notes (e.g., note type, section name, word-level semantic category extracted from UMLS). These handcrafted features usually lead to high precision but are not general enough to recall unseen mentions. In addition, they usually do not generalise well in different types of text (i.e., online posts in CADEC). In contrast, we avoid these handcrafted features in our model.

5.5 Analysis

5.5.1 Impact of mention and interval length

Discontinuous mentions usually represent compositional concepts that consist of multiple components. Therefore, these mentions are usually longer than continuous mentions (Table 5.1). In addition, intervals between components make the total length of span involved even longer. Previous work shows that flat NER performance degrades when applied on long mentions (Augenstein et al., 2017; Xu et al., 2017; Lange et al., 2020).

We experiment to measure the ability of different models on recalling mentions of different lengths, and to observe the impact of interval lengths. We find that the recall of all models decreases with the increase of mention length in general (Figure 5.3 (a – c)), which is similar to previous observations in the literature on flat mentions Lange et al. (2020). However, the impact of interval length is not straightforward. Mentions with very short interval lengths are as difficult as those with very long interval lengths to be recognised (Figure 5.3 (d – f)). On CADEC, discontinuous mentions with interval length of two are easiest to be recognised (Figure 5.3 (d)), whereas those with interval length of three are easiest on SHARE/CLEF 2013 and 2014. We hypothesise this also relates to annotation inconsistency, because very short intervals may be overlooked by annotators.

Our method achieves highest recall among all models in most settings. This demonstrates our model is effective to recognise both continuous and discontinuous mentions with various lengths. In contrast, the BIO extension model is only strong at recalling continuous mentions (outperforming the graph-based model), but fails on discontinuous mentions (interval lengths larger than zero).

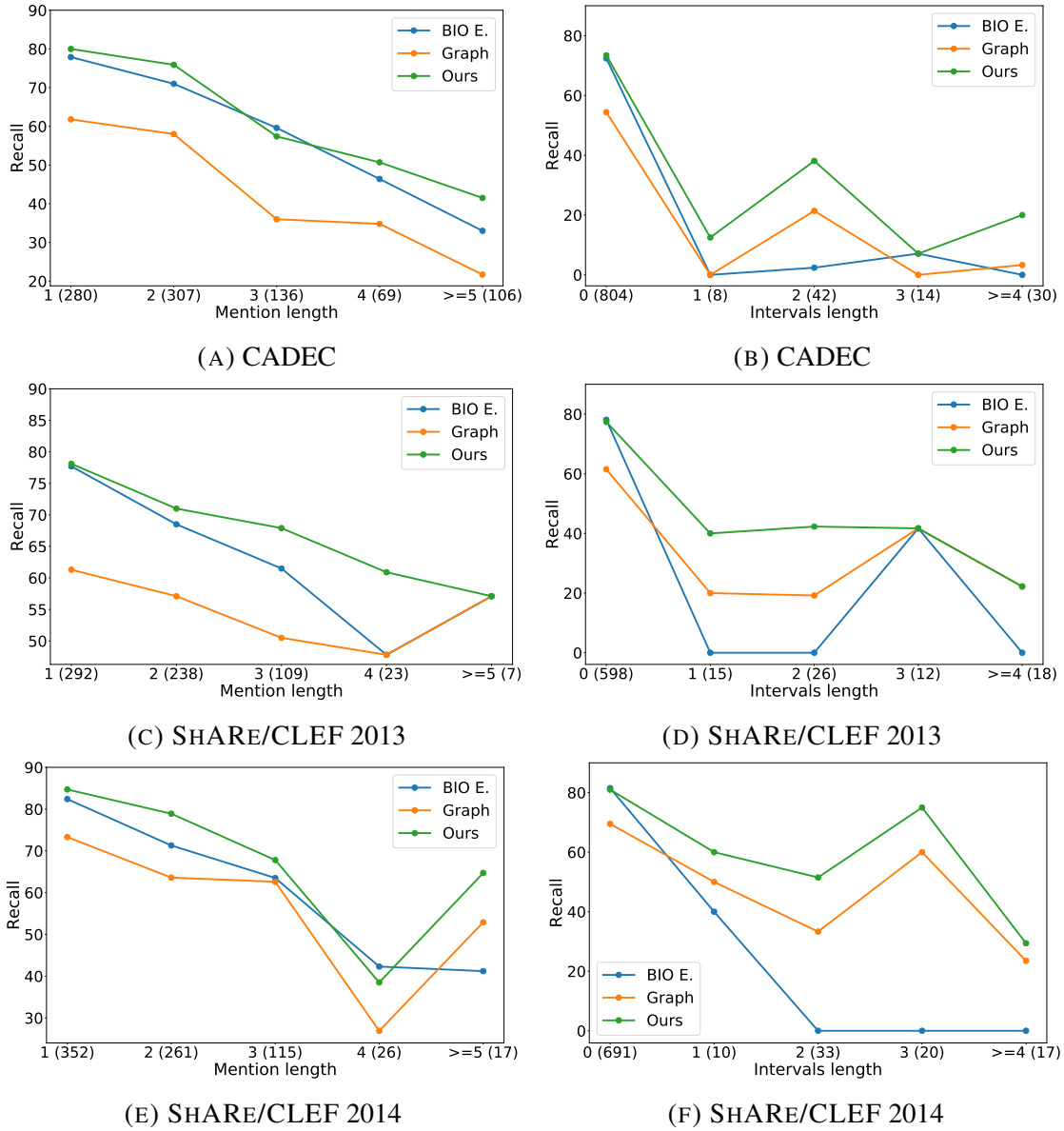


FIGURE 5.3. The impact of mention length and interval length on recall. Mentions with interval length of zero are continuous mentions. Numbers in parentheses are the number of gold mentions.

5.5.2 Impact of overlapping structure

Another characteristic of discontinuous mentions is that they usually *overlap* (Section 5.2). Previous study shows that the intervals between components can be problematic for coordination boundary detection (Ficler and Goldberg (2016)). Conversely, we want to observe

		CADEC		ShARe 2013		ShARe 2014	
		#	F	#	F	#	F
No overlap	BIO extension		0.0		7.5		0.0
	Graph	9	0.0	41	32.1	39	45.2
	Ours		0.0		36.1		57.1
Overlap at left	BIO extension		6.0		25.0		15.7
	Graph	54	9.2	11	45.5	30	37.7
	Ours		28.6		33.3		49.2
Overlap at right	BIO extension		0.0		0.0		0.0
	Graph	16	45.2	19	21.4	5	0.0
	Ours		29.3		13.3		0.0
Multiple overlaps	BIO extension		0.0		–		0.0
	Graph	15	0.0	0	–	6	0.0
	Ours		0.0		–		0.0

TABLE 5.5. Evaluation results on different categories of discontinuous mentions. ‘#’ columns show the number of gold discontinuous mentions in development set of each category.

whether the overlapping structure may help or hinder discontinuous entity recognition. We categorise discontinuous mentions into different subsets, described in Section 5.2, and measure the effectiveness of different discontinuous NER models on each category.

From Table 5.5, we find that our model achieves better results on discontinuous mentions belonging to *No overlap* category on SHARE/CLEF 2013 and 2014, and *Overlap at left* category on CADEC and SHARE/CLEF 2014. Note that *No overlap* category accounts for half of discontinuous mentions in SHARE/CLEF 2013 and 2014, whereas *Overlap at left* accounts for half in CADEC (Table 5.1). Graph-based model achieves better results on *Overlap at right* category. On the *Multiple overlaps* category, no models is effective², which emphasises the challenges of dealing with this syntactic phenomena (examples can be found in Section 5.5.3). We note, however, the portion of discontinuous mentions belonging to this category is very small in all three datasets.

Although our model achieves better results on *No overlap* category on SHARE/CLEF 2013 and 2014, it does not predict correctly any discontinuous mention belonging to this category

²Our model cannot recognise all mentions belonging to this category in theory. For example, if two mentions overlap at both the left and the right, our model can predict only one of them.

on CADEC. The ineffectiveness of our model, as well as other discontinuous NER models, on CADEC *No overlap* category can be attributed to two reasons: 1) the number of discontinuous mentions belonging to this category in CADEC is small (around 12%), rendering the learning process more difficult. 2) the gold annotations belonging to this category are inconsistent from a linguistic perspective. For example, severity indicators are annotated as the interval of the discontinuous mention sometimes, but not often. Note that this may be reasonable from a medical perspective, as some symptoms are roughly grouped together no matter their severity, whereas some symptoms are linked to different concepts based on their severity and severe adverse drug reactions are especially on the radar.

5.5.3 Example predictions

We find that previous models often fail to identify discontinuous mentions that involve long and overlapping spans. For example, the sentence *‘Severe joint pain in the shoulders and knees.’* contains two mentions: *‘Severe joint pain in the shoulders’* and *‘Severe joint pain in the knees’*. Graph-based model does not identify any mention from this sentence, resulting in a low recall. The BIO extension model predicts most of these tags (8 out of 9) correctly, but fails to decode into correct mentions (predict *‘Severe joint pain in the’*, resulting in a false positive, while it misses *‘Severe joint pain in the shoulders’*). In contrast, our model correctly identifies both of these two mentions.

Another observation is that no model can fully recognise mentions which form crossing compositions. For example, the sentence *‘Joint and Muscle Pain / Stiffness’* contains four mentions: *‘Joint Pain’*, *‘Joint Stiffness’*, *‘Muscle Stiffness’* and *‘Muscle Pain’*, all of which share multiple components with the others. Our model correctly predicts *‘Joint Pain’* and *‘Muscle Pain’*, but it mistakenly predicts *‘Stiffness’* itself as a mention (Table 5.6).

Sentence	Walked like that for about six months with increasing pain , especially in right thigh which felt .
Gold mentions	1. pain in right thigh
Predictions	1. pain [BIO extension] No prediction [Graph] 1. increasing pain [Ours]
Sentence	Stated with joint and pain and muscle weakness , depression , fatigue and cramps .
Gold mentions	1. joint pain ; 2. <i>muscle weakness</i> ; 3. <i>depression</i> ; 4. <i>fatigue</i> ; 5. <i>cramps</i>
Predictions	1. joint and pain ; 2. <i>muscle weakness</i> ; 3. <i>depression</i> ; 4. <i>fatigue</i> ; 5. <i>cramps</i> [BIO extension] No prediction [Graph] 1. <i>muscle weakness</i> ; 2. pain weakness ; 3. joint weakness ; 4. <i>depression</i> ; 5. <i>fatigue</i> ; 6. <i>cramps</i> [Ours]
Sentence	stopped taking them 4 years ago and still suffer terrible muscle pain and wasting .
Gold mentions	1. <i>muscle pain</i> ; 2. muscle wasting
Predictions	1. terrible muscle pain ; 2. wasting [BIO extension] No prediction [Graph] 1. <i>muscle pain</i> ; 2. wasting [Ours]
Sentence	Then I sated having hip / leg / foot pain and numbness .
Gold mentions	1. hip pain ; 2. leg pain ; 3. <i>foot pain</i> ; 4. <i>numbness</i>
Predictions	1. <i>foot pain</i> ; 2. pain ; 3. <i>numbness</i> [BIO extension] No prediction [Graph] 1. stated ; 2. <i>hip pain</i> ; 3. <i>leg pain</i> ; 4. <i>foot pain</i> ; 5. <i>numbness</i> [Ours]
Sentence	Severe joint pain in the shoulders and knees .
Gold mentions	1. <i>Severe joint pain in the shoulders</i> ; 2. <i>Severe joint pain in the knees</i>
Predictions	1. Severe joint pain in the ; 2. <i>Severe joint pain in the knees</i> [BIO extension] None [Graph] 1. <i>Severe joint pain in the shoulders</i> ; 2. <i>Severe joint pain in the knees</i> [Ours]
Sentence	Joint and Muscle Pain / Stiffness .
Gold mentions	1. Joint pain ; 2. Muscle Stiffness ; 3. <i>Muscle Pain</i> ; 4. Joint Stiffness
Predictions	1. Joint ; 2. <i>Muscle Pain</i> ; 3. <i>Stiffness</i> [BIO extension] 1. <i>Joint pain</i> ; 2. Muscle Pain / Stiffness ; 3. <i>Stiffness</i> [Graph] 1. <i>Joint pain</i> ; 2. <i>Muscle Pain</i> ; 3. Stiffness [Ours]

TABLE 5.6. Example sentences involving discontinuous entity mentions and predictions using different methods. These examples are taken from CADEC. Gold discontinuous mentions are highlighted in bold. We cross out the incorrect predictions (false positives) for easy understanding.

5.5.4 Ablation studies

To empirically evaluate the importance of attention and ELMo components, we test the performance of model variants where attention and ELMo are removed separately on CADEC and SHARE/CLEF 2013 datasets.

Model	CADEC		SHARE/CLEF 2013	
	All	Subset w. Disc.	All	Subset w. Disc.
Full	68.4	65.4	77.2	64.3
-Attention	68.4	63.3	76.8	62.3
-ELMo	66.7	62.2	75.2	60.9

TABLE 5.7. Ablation study to estimate the contribution of attention and ELMo components.

The results in Table 5.7 show that removing attention hurts the performance when evaluated on sentences with discontinuous mentions (*w. Disc.* columns), but have little impact on the complete test set where continuous mentions are prevalent. Since we use BiLSTM to derive contextual representation for each token, we believe these contextual representations are effective at recognising continuous mentions, but have trouble identifying intervals within discontinuous mentions. Attention mechanism, via allowing tokens interacting with distant tokens, can capture additional discontinuous dependencies which are not captured by BiLSTM. In terms of the ELMo component, we find that it contributes approximately 2 F_1 score when evaluated on the complete test set and around 4 F_1 when evaluated on sentences with discontinuous mentions, demonstrating the usefulness of pretrained word representations.

5.6 Summary

Recognising discontinuous mentions that represent compositional concepts is important for downstream applications such as pharmacovigilance. We propose an end-to-end transition-based model for discontinuous NER. It makes use of specialised actions and attention mechanism to determine whether a span is the component of a discontinuous mention or not. We evaluate our model on three biomedical datasets with a substantial number of discontinuous mentions and demonstrate that our model can effectively recognise discontinuous mentions without sacrificing the accuracy on continuous mentions. Analysis also suggests that our model is better than existing discontinuous NER models at handling long mentions, resulting in higher recall.

Conclusions

Recognising biomedical names from scholarly articles, clinical notes, and social media data is a fundamental NLP task that can benefit many downstream biomedical NLP and information retrieval applications. However, due to the unique characteristics of biomedical names and the stylistic variation in biomedical language—used by biomedical researchers, practitioners, patients and other participants—biomedical NER needs to solve challenges comparatively less studied in the generic domain NER applications.

In this thesis, we first identified challenges of applying standard sequence tagger to recognise biomedical names. Although sequence tagging techniques have demonstrated their effectiveness in generic domain NER, achieving state-of-the-art performance in many benchmarks, they suffer from three problems when being applied in the biomedical domain:

- Biomedical names may consist of non-consecutive spans and they may overlap with each other. The main reason of this complex structure in biomedical names is that many biomedical concepts are compositional. For example, a symptom description may consist of several components: body location, severity indicator, and general feeling, and these components may locate far away from each other.
- Training of neural based sequence taggers usually requires large training set, which is difficult to obtain in the biomedical domain. Annotating biomedical NER datasets usually requires domain-knowledge, and sometimes even unlabelled data are unavailable due to legal reasons.
- State-of-the-art sequence taggers are usually enhanced by language representation models pre-trained on large set of generic domain unlabelled data. Domain shift

between these out-of-domain pre-training data and the target biomedical data usually results in a performance drop.

Targeting these three problems, we explored the corresponding research directions.

We proposed a transition-based model for discontinuous NER. The proposed model is an end-to-end model with generic neural encoding that allows us to leverage specialised actions and attention mechanism to determine whether a span is the component of a discontinuous mention or not. We evaluate our model on three biomedical datasets with a substantial number of discontinuous mentions and demonstrate that our model can effectively recognise discontinuous mentions without sacrificing the accuracy on continuous mentions.

We designed several easy to use data augmentation methods for the NER task: Label-wise token replacement, Synonym replacement, Mention replacement and Shuffle within segments. These augmentations do not rely on any externally trained models, such as machine translation models or syntactic parsing models, which are by themselves difficult to train in a low-resource domain-specific scenario. Through experiments on two biomedical datasets, we show that simple data augmentation can improve performance even over strong baselines, where large scale pre-trained language representation models are used. We leave the exploration of combining these data augmentation methods with other NER models, such as the transition-based model we proposed, for future work.

We analysed different aspects of similarity between domains, and employed cost-effective measures to quantify domain similarity. We demonstrated that these measures are good predictors of the usefulness of pre-trained language representation models on downstream NER task. We find that human intuition favour field (the subject matter being discussed) over tenor (the participants of the discourse and their purpose) when they select in-domain pre-training data. Results suggest that this intuition may be unreliable when the target data set locates in the intersection of several domains.

Based on the discoveries presented in this thesis, we see two future directions worth exploring.

- The first one is on incorporating existing biomedical knowledge base. In Chapter 3, we explored the data augmentation methods that make use of the original training set and a generic lexical database of English. Similar augmentation methods can also be applied to biomedical knowledge base. For example, a biomedical concept—defined using CUI in UMLS—can have several aliases from various vocabularies. These aliases can be used to create augmented sentences in the data augmentation settings. In Chapter 4, we investigated the impact of pre-training data on the effectiveness of pre-trained language representation models, where we pre-train models using only the unlabelled text. We believe more sophisticated pre-training tasks based on biomedical knowledge base can create pre-trained models that capture both language and biomedical knowledge.
- The second direction is on investigating the impact of NER performance on downstream tasks, such as entity linking, relation extraction, or biomedical literature search. In Chapter 5, we showed that our proposed model can effectively recognise discontinuous biomedical names without sacrificing the performance of continuous ones. It worth investing how this improvement can benefit downstream tasks whose results are directly presented to end users.

Bibliography

- Alan Agresti. 2003. *Categorical data analysis*. John Wiley & Sons.
- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico.
- Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. In *Biological, translational, and clinical language processing*, pages 65–72, Prague, Czech Republic.
- Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota.
- Nora Anderson and Juergen Borlak. 2011. Correlation versus causation? pharmacovigilance of the analgesic flupirtine exemplifies the need for refined spontaneous ADR reporting. *PloS one*, page e25221.
- Shlomo Argamon-Engelson and Ido Dagan. 1999. Committee-based sample selection for probabilistic classifiers. *Journal of Artificial Intelligence Research*, pages 335–360.
- Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17, Washington, DC.
- Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, pages 229–236.
- Vincent Van Asch and Walter Daelemans. 2010. Using domain similarity for performance estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 31–36, Uppsala, Sweden.

- Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada.
- Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5360–5369, Hong Kong, China.
- Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, pages 267–292.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China.
- Emily M. Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, pages 135–146.
- Dorothy Bonn. 1998. Adverse drug reactions remain a major cause of death. *The Lancet*, page 1183.
- Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal.

- John S Brownstein, Shawn N Murphy, Allison B Goldfine, Richard W Grant, Margarita Sordo, Vivian Gainer, Judith A Colecchi, Anil Dubey, David M Nathan, and John P Glaser. 2010. Rapid identification of myocardial infarction risk associated with diabetes medications using electronic medical records. *Diabetes care*, pages 526–531.
- Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6334–6343, Online.
- David Campos, Sérgio Matos, and José Luís Oliveira. 2012. *Biomedical named entity recognition: a survey of machine-learning tools*. IntechOpen.
- Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.
- Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th workshop on biomedical natural language processing*, pages 166–174, Berlin, Germany.
- Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2089–2095, Minneapolis, Minnesota.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, Online.
- Maximin Coavoux and Shay B Cohen. 2019. Discontinuous constituency parsing with a stack-free transition system and a dynamic oracle. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 204–217, Minneapolis, Minnesota.

- Maximin Coavoux, Benoît Crabbé, and Shay B Cohen. 2019. Unlexicalized transition-based discontinuous constituency parsing. *Transactions of the Association for Computational Linguistics*, pages 73–89.
- Nigel Collier, Hyun Seok Park, Norihiro Ogata, Yuka Tateisi, Chikashi Nobata, Tomoko Ohta, Tateshi Sekimizu, Hisao Imai, Katsutoshi Ibushi, and Jun'ichi Tsujii. 1999. The genia project: corpus-based knowledge acquisition and information extraction from genome research papers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–272, Bergen, Norway.
- Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110, College Park, MD.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, pages 2493–2537.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online.
- Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116, Valencia, Spain.
- Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *Proceedings of the Twentieth National Conference on Artificial Intelligence*, pages 746–751, Pittsburgh, Pennsylvania.
- Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087, Montreal, Canada.
- Xiang Dai. 2018. Recognizing complex entity mentions: A review and future directions. In *Proceedings of ACL 2018, Student Research Workshop*, pages 37–44, Melbourne, Australia.

- Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online).
- Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2019. Using similarity measures to select pretraining data for NER. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1460–1470, Minneapolis, Minnesota.
- Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020a. Cost-effective selection of pretraining data: A case study of pretraining BERT on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1675–1681, Online.
- Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020b. An effective transition-based model for discontinuous NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5860–5870, Online.
- Xiang Dai, Sarvnaz Karimi, and Cecile Paris. 2017. Medication and adverse event extraction from noisy text. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 79–87, Brisbane, Australia.
- Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.
- Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, pages 1–10.

- Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*, Toulon, France.
- Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343, Beijing, China.
- Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, pages 91–134.
- Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada.
- Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource pos tagging using cross-lingual projection. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 178–186, Berlin, Germany.
- Joshua C Feblowitz, Adam Wright, Hardeep Singh, Lipika Samal, and Dean F Sittig. 2011. Summarization of clinical information: a conceptual model. *Journal of biomedical informatics*, pages 688–699.
- Jessica Fidler and Yoav Goldberg. 2016. A neural network for coordination boundary prediction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 23–32, Austin, Texas.
- Jenny Rose Finkel and Christopher D Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 141–150, Singapore.
- Radu Florian, Hany Hassan, Abraham Ittycheriah, Hongyan Jing, Nanda Kambhatla, Xiaoqi-ang Luo, H Nicolov, and Salim Roukos. 2004. A statistical model for multilingual entity

- detection and tracking. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts.
- Bent Fuglede and Flemming Topsøe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.*, page 31.
- Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544, Florence, Italy.
- Alex Graves, Abdel rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649, Vancouver, Canada.
- Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online.
- Muhammad Abdul Hadi, Chin Fen Neoh, Rosdi M Zin, Mahmoud E Elrggal, and Ejaz Cheema. 2017. Pharmacovigilance: pharmacists’ perspective on spontaneous adverse drug reaction reporting. *Integrated pharmacy research and practice*, page 91.
- M. A. K. Halliday and Ruqaiya Hasan. 1989. *Language, Context and Text: Aspects of Language in a Social-Semiotic Perspective*. Deakin University Press.
- Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pages 8527–8537, Montréal, Canada.

- Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China.
- Lorna Hazell and Saad AW Shakir. 2006. Under-reporting of adverse drug reactions. *Drug safety*, pages 385–396.
- Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197, Edinburgh, Scotland.
- Michael A Hedderich and Dietrich Klakow. 2018. Training a neural network in a low-resource setting on automatically annotated noisy data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 12–18, Melbourne, Australia.
- Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv:2010.12309*.
- Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2005. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*.
- Jerry R Hobbs. 2002. Information extraction from biomedical text. *Journal of biomedical informatics*, pages 260–264.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, pages 1735–1780.
- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 541–550, Portland, Oregon.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia.

- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, pages 1–9.
- Mark Johnson, Peter Anderson, Mark Dras, and Mark Steedman. 2018. Predicting accuracy on large datasets from smaller pilot data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 450–455, Melbourne, Australia.
- Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C Raina MacIntyre. 2019. Survey of text-based epidemic intelligence: A computational linguistics perspective. *ACM Computing Surveys*, pages 1–19.
- Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1446–1459, New Orleans, Louisiana.
- Sarvnaz Karimi, Xiang Dai, Hamed Hassanzadeh, and Anthony Nguyen. 2017. Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. In *Proceedings of the 16th BioNLP Workshop*, pages 328–332, Vancouver, Canada.
- Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015a. CA-DEC: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, pages 73–81.
- Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015b. Text and data mining techniques in adverse drug reaction detection. *ACM Computing*

- Surveys (CSUR)*, pages 1–39.
- Arzoo Katiyar and Claire Cardie. 2018. Nested named entity recognition revisited. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 861–871, New Orleans, Louisiana.
- H Khalil and C Huang. 2020. Adverse drug reactions in primary care: a scoping review. *BMC Health Services Research*, page 5.
- J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun’ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, pages i180–i182.
- Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, pages 70–75, Geneva, Switzerland.
- Dietrich Klakow. 2000. Selecting articles from the language model training corpus. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1695–1698, Istanbul, Turkey.
- Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 817–824, Sydney, Australia.
- Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana.
- Isaac S Kohane, Daniel R Masys, and Russ B Altman. 2006. The incidentalome: a threat to genomic medicine. *The Journal of the American Medical Association*, pages 212–215.
- Chaitanya Kulkarni, Wei Xu, Alan Ritter, and Raghu Machiraju. 2018. An annotated corpus for machine reading of instructions in wet lab protocols. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 97–106, New Orleans,

Louisiana.

- Michal Laclavík, Štefan Dlugolinský, Martin Šeleng, Marcel Kvassay, Emil Gatial, Zoltán Balogh, and Ladislav Hluchý. 2012. Email analysis and information extraction for enterprise benefit. *Computing and informatics*, pages 57–87.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California.
- Lukas Lange, Xiang Dai, Heike Adel, and Jannik Strötgen. 2020. NLNDE at CANTEMIST: Neural sequence labeling and parsing approaches for clinical concept extraction. In *Iberian Languages Evaluation Forum (IberLEF 2020)*, pages 335–346, Online.
- Lukas Lange, Michael A Hedderich, and Dietrich Klakow. 2019. Feature-dependent confusion matrices for low-resource ner labeling with noisy labels. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3554–3559, Hong Kong, China.
- Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, pages 28–37.
- Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts in health-related social networks. In *Proceedings of the 2010 workshop on biomedical natural language processing*, pages 117–125, Uppsala, Sweden.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, pages 436–444.
- David YW Lee. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the bnc jungle. *Language Learning and Technology*, pages 37–72.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model

- for biomedical text mining. *Bioinformatics*, pages 1234–1240.
- Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.
- Peng-Hsuan Li, Ruo-Ping Dong, Yu-Siang Wang, Ju-Chieh Chou, and Wei-Yun Ma. 2017. Leveraging linguistic structures for named entity recognition with bidirectional recursive neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2664–2669, Copenhagen, Denmark.
- Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1054–1064, Online.
- Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2019. Sequence-to-nuggets: Nested entity mention detection via anchor-region networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5182–5192, Florence, Italy.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 857–867, Lisbon, Portugal.
- Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium.
- Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3036–3046, Minneapolis, Minnesota.

- Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal.
- Xuezhong Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany.
- Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Flexible text segmentation with structured multilabel classification. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 987–994, Vancouver, Canada.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.
- Sepideh Mesbah, Jie Yang, Robert-Jan Sips, Manuel Valle Torre, Christoph Lofi, Alessandro Bozzon, and Geert-Jan Houben. 2019. Training data augmentation for detecting adverse drug reactions in user-generated content. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2349–2359, Hong Kong, China.
- Alejandro Metke-Jimenez and Sarvnaz Karimi. 2016. Concept identification and normalisation for adverse drug event discovery in medical forums. In *Proceedings of the First International Workshop on Biomedical Data Integration and Discovery*, Kobe, Japan.
- Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8, Bergen, Norway.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhresch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 52–55, Miyazaki, Japan.

- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *27th Annual Conference on Neural Information Processing Systems*, pages 3111–3119, Lake Tahoe, Nevada.
- George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, pages 235–244.
- Junghyun Min, R Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online.
- Robert C Moore and Will Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden.
- Johannes P Mouton, Ushma Mehta, Andy G Parrish, Douglas PK Wilson, Annemie Stewart, Christine W Njuguna, Nicole Kramer, Gary Maartens, Marc Blockman, and Karen Cohen. 2015. Mortality from adverse drug reactions in adult medical inpatients at four hospitals in south africa: a cross-sectional survey. *British journal of clinical pharmacology*, pages 818–826.
- Danielle L Mowery, Sumithra Velupillai, Brett R South, Lee Christensen, David Martinez, Liadh Kelly, Lorraine Goeriot, Noemie Elhadad, Sameer Pradhan, and Guergana Savova. 2014. Task 2: ShARe/CLEF ehealth evaluation lab 2014. In *Conference and Labs of the Evaluation Forum*, pages 31–42, Sheffield, United Kingdom.
- Aldrian Obaja Muis and Wei Lu. 2016. Learning to recognize discontinuous entities. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 75–84, Austin, Texas.
- Aldrian Obaja Muis and Wei Lu. 2017. Labeling gaps between words: Recognizing overlapping mentions with mention separators. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2608–2618, Copenhagen, Denmark.

- n2c2. 2019. [Track 2: n2c2/OHNL track on family history extraction](#). Web page, Harvard Medical School – Department of Biomedical Informatics.
- David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, pages 3–26.
- Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online.
- Joel Nothman, James R Curran, and Tara Murphy. 2008. Transforming wikipedia into named entity training data. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, pages 124–132, Hobart, Australia.
- Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia.
- Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing. Technical report, Swedish Institute of Computer Science.
- Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1345–1359.
- Barun Patra and Joel Ruben Antony Moniz. 2019. Weakly supervised attention networks for entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6269–6274, Hong Kong, China.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, Doha, Qatar.
- Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada.

- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana.
- Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia.
- Toni D. Piazza-Hepp and Dianne L. Kennedy. 1995. Reporting of adverse events to medwatch. *American Journal of Health-System Pharmacy*, pages 1436–1439.
- Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA.
- Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Lee M Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana K Savova. 2013. Task 1: ShARe/CLEF ehealth evaluation lab 2013. In *Conference and Labs of the Evaluation Forum*, pages 212–31, Valencia, Spain.
- Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy.
- Jonathan Raiman and John Miller. 2017. Globally normalized reader. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1059–1069, Copenhagen, Denmark.
- Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online).
- Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R Curran. 2019. NNE: A dataset for nested named entity recognition in English newswire. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,

- pages 5176–5181, Florence, Italy.
- Nicola Ringland. 2016. *Structured Named Entities*. Thesis, University of Sydney.
- Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the gap: Attending to discontinuity in identification of multiword expressions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2692–2698, Minneapolis, Minnesota.
- Sebastian Ruder. 2019. *Neural transfer learning for natural language processing*. Thesis, National University of Ireland Galway.
- Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019a. Latent multi-task architecture learning. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 4822–4829, Honolulu, Hawaii.
- Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark.
- Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019b. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, pages 569–631.
- Maciej Rybinski, Xiang Dai, Sonit Singh, Sarvnaz Karimi, Anthony Nguyen, et al. 2021. Extracting family history information from electronic health records: Natural language processing analysis. *JMIR Medical Informatics*, 9(4):e24020.
- Esteban Safranchik, Shiyong Luo, and Stephen H Bach. 2020. Weakly supervised sequence tagging from noisy rules. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 5570–5578, New York, NY.
- Gözde Gül Şahin and Mark Steedman. 2019. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009, Brussels, Belgium.
- Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, Stroudsburg, PA.

- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, Edmonton, Canada.
- Sunita Sarawagi. 2008. *Information extraction*. Now Publishers Inc.
- Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O’Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*, pages 202–212.
- Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318, Cascais, Portugal.
- Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2019. Green AI. *arXiv preprint arXiv:1907.10597*.
- Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Cambridge University Press Cambridge.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.
- Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison.
- Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, Honolulu, Hawaii.
- Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, pages 443–460.
- Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep active learning for named entity recognition. In *International Conference on Learning Representations*, Vancouver, Canada.

- Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, page 60.
- Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, and Kuzman Ganchev. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, page S2.
- Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849, Brussels, Belgium.
- Guillermo Solano-Flores. 2006. Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of english language learners. *Teachers College Record*, page 2354.
- Gabriel Stanovsky, Daniel Gruhl, and Pablo Mendes. 2017. Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 142–151, Valencia, Spain.
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France.
- Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine De Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pages 138–144, Osaka, Japan.
- Lin Sun, Fule Ji, Kai Zhang, and Chi Wang. 2019. Multilayer toi detection approach for nested ner. *IEEE Access*, pages 186600–186608.
- Charles Sutton and Andrew McCallum. 2007. *An introduction to conditional random fields for relational learning*. The MIT Press.
- Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany.

- Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu. 2013a. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. *BMC medical informatics and decision making*, page S1.
- Buzhou Tang, Jianguo Hu, Xiaolong Wang, and Qingcai Chen. 2018. Recognizing continuous and discontinuous adverse drug reaction mentions from social media using lstm-crf. *Wireless Communications and Mobile Computing*.
- Buzhou Tang, Yonghui Wu, Min Jiang, Joshua C Denny, and Hua Xu. 2013b. Recognizing and encoding disorder concepts in clinical text using machine learning and vector space model. In *Conference and Labs of the Evaluation Forum*, Valencia, Spain.
- Alexander Tkachenko, Timo Petmanson, and Sven Laur. 2013. Named entity recognition in Estonian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 78–83, Sofia, Bulgaria.
- Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, pages 552–556.
- A Vallano, G Cereza, C Pedròs, A Agustí, I Danés, C Aguilera, and JM Arnau. 2005. Obstacles and solutions for spontaneous reporting of adverse drug reactions in the hospital. *British journal of clinical pharmacology*, pages 653–658.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Long Beach, California.
- Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy.
- Bailin Wang and Wei Lu. 2018. Neural segmental hypergraphs for overlapping mention recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 204–214, Brussels, Belgium.

- Bailin Wang and Wei Lu. 2019. Combining spans into entities: A neural two-stage approach for recognizing discontinuous entities. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6216–6224, Hong Kong, China.
- William Yang Wang and Diyi Yang. 2015. That’s so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2557–2563, Lisbon, Portugal.
- Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018a. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium.
- Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, and Sunghwan Sohn. 2018b. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, pages 34–49.
- Taro Watanabe and Eiichiro Sumita. 2015. Transition-based neural constituent parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1169–1179, Beijing, China.
- Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China.
- Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. *OntoNotes: A large training corpus for enhanced processing*. Springer.
- Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, page 9.

- WHO. 2020. **Safety of medicines – adverse drug reactions**. Web page, World Health Organization.
- Congying Xia, Chenwei Zhang, Tao Yang, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, and Philip Yu. 2019a. Multi-grained named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1430–1440, Florence, Italy.
- Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019b. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy.
- Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.
- Mingbin Xu, Hui Jiang, and Sedtawut Watcharawittayakul. 2017. A local detection approach for named entity recognition and mention detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1247, Vancouver, Canada.
- Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico.
- Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889, Santa Fe, New Mexico.
- Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In *International Conference on Learning Representations (ICLR 2017)*, Toulon, France.
- Kang Min Yoo, Youhyun Shin, and Sang goo Lee. 2019. Data augmentation for spoken language understanding via joint variational generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 7402–7409, Honolulu, Hawaii.
- Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention

- for reading comprehension. In *International Conference on Learning Representations*, Vancouver, Canada.
- Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online.
- Shaodian Zhang and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, pages 1088–1098.
- Shiliang Zhang, Hui Jiang, Mingbin Xu, Junfeng Hou, and Li-Rong Dai. 2015a. The fixed-size ordinally-forgetting encoding method for neural network language models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 495–500, Beijing, China.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015b. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, Montreal, Canada.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana.
- Changmeng Zheng, Yi Cai, Jingyun Xu, Ho fung Leung, and Guandong Xu. 2019. A boundary-aware neural model for nested named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 357–366, Hong Kong, China.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, Santiago, Chile.